

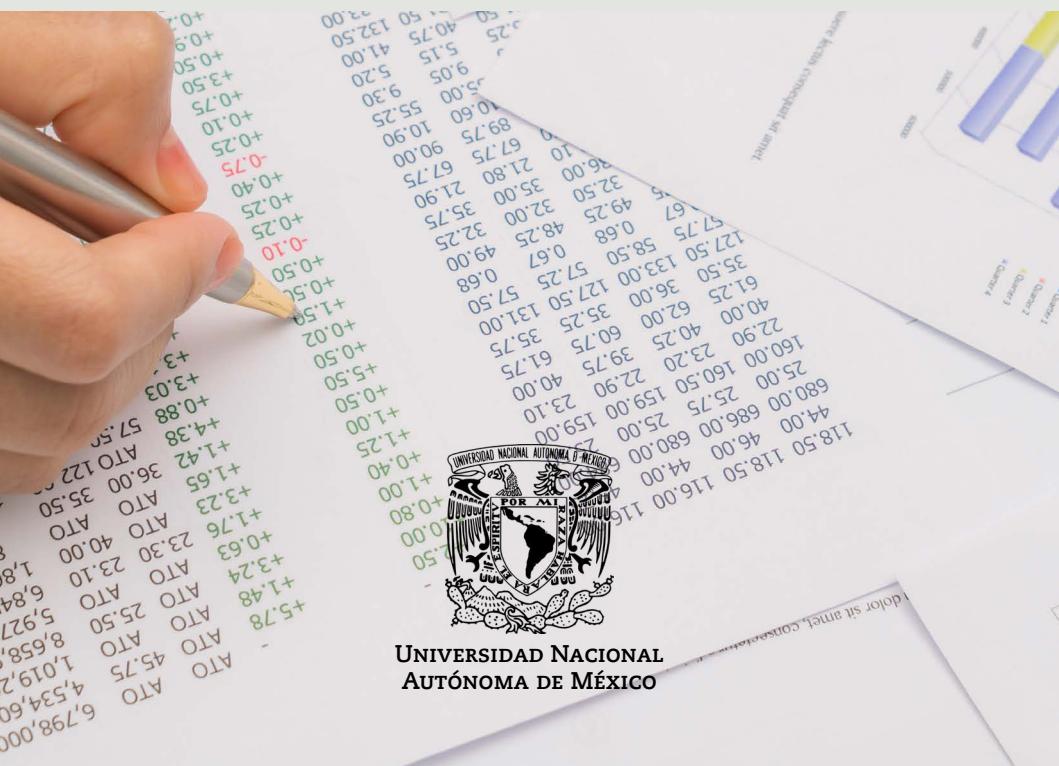
# ¿Cómo empezar a estudiar el mercado de trabajo en México?

Una introducción al análisis estadístico  
con R aplicado a  
la Encuesta Nacional de Ocupación y Empleo

Ana Ruth Escoto Castillo



UNIVERSIDAD NACIONAL  
AUTÓNOMA DE MÉXICO



**Universidad Nacional Autónoma de México**

Enrique Luis Graue Wiechers  
Rector

Leonardo Lomelí Vanegas  
Secretario General

Luis Agustín Álvarez Icaza Longoria  
Secretario Administrativo

Alfredo Sánchez Castañeda  
Abogado General

Socorro Venegas Pérez  
Directora General de Publicaciones y Fomento Editorial

**Facultad de Ciencias Políticas y Sociales**

Carola García Calderón  
Directora

Patricia Guadalupe Martínez Torreblanca  
Secretaria General

Juan Manuel López Ramírez  
Secretario Administrativo

Elvira Teresa Blanco Moreno  
Jefa del Departamento de Publicaciones



**fcps**

# ¿Cómo empezar a estudiar el mercado de trabajo en México?

Una introducción al análisis estadístico  
con *R* aplicado a la  
Encuesta Nacional de Ocupación y Empleo

Ana Ruth Escoto Castillo



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO  
México, 2021

Esta investigación, arbitrada a “doble ciego” por especialistas en la materia, se privilegia con el aval de la Facultad de Ciencias Políticas y Sociales, Universidad Nacional Autónoma de México.

Este libro fue financiado con recursos de la Dirección General de Asuntos del Personal Académico (DGAPA) de la Universidad Nacional Autónoma de México (UNAM), como parte del Programa de Apoyo a Proyectos para Innovar y Mejorar la Educación (PAPIME) PE311019 “Magnitud y características de los procesos laborales. Una aplicación para fortalecer las capacidades de análisis estadístico de los estudiantes de la FCPyS” del que es responsable académica Ana Ruth Escoto Castillo.

**¿Cómo empezar a estudiar el mercado de trabajo en México?**

**Una introducción al análisis estadístico con R aplicado a la  
Encuesta Nacional de Ocupación y Empleo**

Ana Ruth Escoto Castillo

Primera edición: 24 de agosto de 2021

Reservados todos los derechos conforme a la ley.

D.R. © 2021 Universidad Nacional Autónoma de México

Ciudad Universitaria, Alcaldía Coyoacán, C. P. 04510, CDMX, México.

Facultad de Ciencias Políticas y Sociales,

Circuito Mario de la Cueva s/n, Ciudad Universitaria,

Alcaldía Coyoacán, C. P. 04510, CDMX, México.

ISBN: 978-607-30-4913-9

Queda prohibida la reproducción parcial o total, directa o indirecta, del contenido de la presente obra, sin contar previamente con la autorización expresa y por escrito de los editores, en términos de lo así previsto por la Ley Federal de Derechos de Autor y, en su caso, por los tratados internacionales aplicables.

Las opiniones y los contenidos incluidos en esta publicación son responsabilidad exclusiva del/los autor/es.

Libro electrónico hecho en México/e-book made in Mexico

# Índice

<b>Agradecimientos</b>	9
<b>Introducción</b>	11
<b>I. La Encuesta Nacional de Ocupación y Empleo (ENOE): sus temas y aplicaciones</b>	15
Introducción	15
A. Historia de la Encuesta Nacional de Ocupación y Empleo (ENOE)	15
B. Diseño y elementos básicos	17
C. Cobertura temática	19
D. Principales cambios en el tiempo	22
E. Reseña de los estudios que usan la Encuesta Nacional de Ocupación y Empleo en esta década	24
a. Numeralia	25
b. Los temas de la última década	26
c. Desde dónde y cómo se estudia	31
<b>II. Una muy breve introducción a R y preparación de la fuente de información</b>	35
Introducción	35
A. Paquetes utilizados	36
B. Elementos básicos	38
a. La paquetería de R	39
b. ¿Dónde estamos trabajando?	41
c. Comandos básicos para empezar	41
d. Objetos	42
C. Importando la Encuesta a un ambiente de R	43
D. Fusión de las bases de datos	45
a. Fusionado uno a uno	46
b. Tipos de fusionado según información de las bases	49
i) Casos comunes en las dos bases	49
ii) Todos los casos en ambas bases	50

<i>iii) Casos de la base 1</i>	50
<i>iv) Casos de la base 2</i>	51
c. Fusionando bases de diferente nivel	52
E. Revisión breve de la ENOE	53
F. Selección de casos y de variables	54
a. Selección de variables	54
b. Selección “inversa”	55
c. Subconjuntos de datos	55
d. Uso de etiquetas importadas y cómo usarlas	56
<b>III. El análisis descriptivo como herramienta de análisis en los mercados de trabajo. El caso de las variables cualitativas</b>	59
Introducción	59
A. Tipos de variables y escalas de medición	60
a. Variables nominales	62
b. Variables ordinales	64
B. Tablas de doble entrada	67
a. Cálculo de frecuencias	67
b. Totales y porcentajes	68
C. Gráficas de barra	70
a. Barras de una variable	71
b. Gráficas de barra de dos variables cualitativas	78
c. Sobre las paletas de colores	81
D. Ejemplo a aplicación: estructura productiva y sexo	84
E. Las tasas de participación económica	89
F. Análisis descriptivo. La utilidad de los datos ponderados	99
a. Replicando los datos del INEGI	100
<b>IV. El análisis descriptivo de las variables numéricas. El caso de los ingresos laborales en la ENOE</b>	104
Introducción	104
A. Análisis descriptivo univariado	105
a. Las primeras gráficas	106
<i>i) Gráfica de tallo y hoja</i>	107
<i>ii) Histograma</i>	108
<i>iii) Gráfica de densidad</i>	109
b. Un atajo <i>esquisse</i>	111
c. Las medidas numéricas: la media y la desviación estándar	114
d. El resumen de cinco números y las gráficas de caja y brazos	116
B. Los ingresos laborales en la ENOE	120

C. Comparando ingresos entre grupos	125
a. Gráficas para mostrar diferentes categorías	125
b. Estadísticos para grupos	127
c. Estadísticos con datos expandidos	128
D. El coeficiente de Gini	130
E. La relación entre dos variables cuantitativas	133
a. Gráfica de dispersión	133
b. Correlación	137
 <b>V. Introducción a la inferencia</b>	 139
Introducción	139
A. La media poblacional $\mu$	143
B. Diferencia de medias de dos grupos	147
C. Estimación de varianzas y sus pruebas de hipótesis	150
D. Estimación de diferencias de varianzas y sus pruebas de hipótesis	154
E. Prueba chi-cuadrado: una aplicación para inferencia en tablas de contingencia	157
F. Análisis de la varianza de un factor: comparación de varias medias	159
a. Apreciación gráfica	160
b. Comparación entre grupos	162
c. Supuestos del análisis de varianza	163
i) <i>Prueba de normalidad</i>	163
ii) <i>Prueba de Bartlett para homogeneidad de varianzas</i>	164
d. ¿Qué hacer? Un método: la prueba <i>Kruskal-Wallis test</i>	165
 <b>VI. La regresión lineal. Un ejemplo para modelar los ingresos en los mercados laborales</b>	 167
Introducción	167
A. La línea de mínimos cuadrados ordinarios "(MCO)"	168
B. Regresión lineal simple	171
a. Aplicación de una regresión lineal simple	174
b. Supuestos y su diagnóstico	177
C. Regresión lineal múltiple	182
a. Agregando una variable categórica	182
b. Un modelo más complejo	185
D. Presentación de modelos	186
E. Estandarizando las unidades de medida las variables	189
F. No cumple los supuestos ¿Y ahora qué?	191
a. Heterocedasticidad	191
b. Regresión robusta a valores atípicos	192

<b>VII. Aplicaciones longitudinales con la ENOE.</b>	
<b>El caso del análisis de secuencias</b>	196
Introducción	196
A. Construcción del panel de la ENOE	197
B. Formato largo vs. formato ancho	202
C. El análisis de secuencias. Una introducción	205
a. Gráficas aluviales	207
b. Análisis de secuencias con el paquete <i>TrMineR</i>	209
c. Descripción de las secuencias	210
D. Tasas de transición	215
<b>Palabras finales</b>	220
<b>Referencias</b>	223
<b>Glosario</b>	233
<b>Anexo: Análisis de texto de las investigaciones en México</b>	238
Bibliografía analizada	238
Código de análisis	255

## Agradecimientos

Este libro es resultado del proyecto “Magnitud y características de los procesos laborales. Una aplicación para fortalecer las capacidades de análisis estadístico de los estudiantes de la FCPyS”, financiado por la Dirección General de Asuntos del Personal Académico de la UNAM mediante el Programa de Apoyo a Proyectos para la Innovación y Mejoramiento de la Enseñanza (PAPIME). Agradezco el apoyo y financiamiento de nuestra institución para el desarrollo del proyecto y de este texto.

Mi más amplio reconocimiento a la labor de Fernanda Morfín y Elsie Carolina López, becaria y participante del proyecto, respectivamente, sin quienes este texto no hubiera salido a la luz. Les agradezco su apoyo logístico y para la revisión bibliográfica.

Del mismo modo, agradezco la lectura y comentarios a Mónica Lara Escalante en la parte técnica y a Rodrigo Hernández en la parte gramatical y de lenguaje. Seis ojos son mejores que dos.

Una mención especial y agradecimiento a Nina Castro Méndez por haberme inspirado a utilizar más *R* y haberme acompañado en talleres y en el desarrollo de los códigos.

Por supuesto, también agradezco a la comunidad estudiantil. Sin ellos y ellas este libro no tendría razón de existir.

Ciudad Universitaria, Coyoacán, 2021.

*A Nina, Edith, Emmalí, Mauri, Paty, Viri, Isalia y Nelson.*

*Gracias por enseñarme y acompañarme.*

# Introducción

México es un país desigual. Una de las desigualdades que apremian más a la población es la que refiere a los ingresos. En un país altamente mercantilizado, el poder de compra es el medio principal para acceder a los bienes y servicios que nos proveen bienestar. Los ingresos por trabajo representan más del 70% de todos los ingresos, según datos de la Encuesta Nacional de Ingreso y Gasto de los Hogares (ENIGH) de 2018 (INEGI, 2018). Comprender cómo se entraman las desigualdades en el mercado de trabajo es esencial para estudiar uno de los grandes problemas en México.

Con esta idea en mente, en el proyecto PAPIME “Magnitud y características de los procesos laborales. Una aplicación para fortalecer las capacidades de análisis estadístico de los estudiantes de la FCPys” nos propusimos acercar a los y las estudiantes de la Facultad de Ciencias Políticas y Sociales (FCPys) de la Universidad Nacional Autónoma de México (UNAM) a investigaciones cuantitativas sobre el mercado de trabajo.

Este libro pretende ser una herramienta más para este acercamiento, no sólo de los estudiantes de la FCPys, sino de cualquiera que le interese estudiar el mercado de trabajo en México a partir de la Encuesta Nacional de Ocupación y Empleo (ENOE).

En este sentido, el libro es un ejemplo de investigación aplicada y de cómo el análisis empírico de la realidad puede acercar al aprendizaje de técnicas estadísticas. De ahí que podemos retomar algunos elementos de evaluación que se proponen en el campo de las estadísticas “auténticas” aplicadas a otras disciplinas (Hulsizer & Woolf, 2009). La idea es que los procesos de aprendizaje se vuelven más auténticos en la medida que los estudiantes se comprometen a ellos, ya que resuelven un problema particular que les compete.

Es decir, aquí se presenta una forma de cómo se puede estudiar la realidad con datos recolectados desde fuentes oficiales. A partir de ello, se pueden plantear estrategias hacia la evaluación formativa; en

otras palabras, un proceso interactivo mediante el cual los maestros y los estudiantes evalúan lo que han aprendido.

Por tanto, éste no es un libro de estadística ni tampoco de programación; más bien es un ejemplo de cómo se puede comprender la realidad social a partir de la información de las encuestas oficiales en México gracias a técnicas estadísticas, y cómo a partir de ello se pueden plantear estrategias hacia la evaluación formativa; en otras palabras, es un proceso interactivo mediante el cual la población docente y estudiantil evalúan lo que han aprendido.

En específico se utiliza la Encuesta Nacional de Ocupación y Empleo (ENOE), del Instituto Nacional de Estadística y Geografía de México (INEGI), y se utiliza la información provista sobre los trimestres IV de 2018 a IV de 2019. Esta fuente, además de tocar un tema tan importante para la realidad nacional, es consistente a lo largo del tiempo. De ahí que esta guía puede ser utilizada tanto para ejercicios anteriores como posteriores.

Un objetivo de este libro es que quien lo lea no tenga que estar yendo constantemente a revisar los apuntes de estadística o un manual de programación, por lo que se dan descripciones básicas de las técnicas, pero quien las utilice después deberá profundizar. Se utiliza el programa *R* (R Core Team, 2019), que es un programa libre desarrollado con aportes colectivos por una comunidad y que es de libre acceso, lo que supone una barrera mucho menor. En el capítulo II se establece una muy breve introducción a los elementos básicos.

Los códigos que se vierten en esta obra provienen de un largo periodo de investigación en el mercado de trabajo. Si bien su autora ha intentado irlos actualizando y mejorando, seguramente no serán los códigos más eficientes, pero sí han cumplido el objetivo de describir la realidad social. El ejercicio de comentar su aplicación y sus resultados también ha sido una experiencia enriquecedora para quien escribe el texto y ha exigido organizar muchos años de trabajo.

En general, los comandos mantienen comentarios para que quien los utilice entienda cuáles elementos se ponen en los argumentos; siguiendo la lógica “qué me hubiera gustado que me explicaran cuando aprendí a usar esta función”. No es un camino único y comprende lo aprendido en sitios en internet, en los códigos compartidos de los colegas y demás conocimiento comunitario que es difícil de citar.

De ahí que la voz que se use en este libro, sobre todo en la parte más práctica, sea en primera persona colectiva, para rescatar este sentimiento de “nosotros” colectivo para hacer y aprender juntos.

Este libro intenta también abonar al conocimiento colectivo de los mercados de trabajo, en específico el mexicano. Para ello, está estructurado como se expone a continuación.

En el capítulo I se presenta la encuesta, su historia y sus temas. Se hace una breve revisión de los trabajos que han utilizado la encuesta en la presente década, es decir, de 2010 en adelante. También se muestra un breve análisis de los temas y de las técnicas que se han utilizado. El análisis de este capítulo se realizó con *R* y el código utilizado se puede encontrar en el anexo.

En el capítulo II se describe el paquete estadístico utilizado, *R*; pero también se hace un recuento de otros paquetes que el usuario debe instalar, así como sus requerimientos. En la segunda parte del capítulo establecemos un primer acercamiento al manejo de la ENOE. El lector o lectora podrá encontrar una sección de código que permite fusionar la base de la encuesta en su totalidad, ya que está compuesta de varias tablas con unidades diferentes de información.

A partir del capítulo III los textos se vuelven prácticos. Todos incluyen una introducción a las técnicas y se van comentando los resultados que, ya de por sí, son hallazgos de investigación, puesto que provienen del análisis de datos de la encuesta oficial. En la conclusión de cada capítulo quien lee puede encontrar un resumen de estos hallazgos y las preguntas que fueron contestadas con las técnicas.

En el capítulo III se presenta una introducción al análisis descriptivo concentrándose en el análisis de variables de tipo cualitativo. Se muestran dos ejemplos fundamentales para entender el funcionamiento del mercado de trabajo: la estructura por sectores de actividad y las tasas de participación.

En el capítulo IV se continúa con el aproximamiento descriptivo, pero se introducen las particularidades que provienen de estudiar una variable cuantitativa; en este caso, los ingresos laborales por hora. Además, se incluye una introducción a cómo se calcula y se interpreta el índice de Gini, un indicador muy útil para medir desigualdades que, como se estudia en el capítulo I, resulta ser un tema recurrente en los análisis del mercado laboral.

En el capítulo v se plantea una breve introducción a la inferencia, sobre todo para comprender algunos elementos de pruebas de especificación. Cuenta con ejemplos aplicados a proporciones y medias de las variables de la base de datos.

Los capítulos III, IV y V no requieren de muchos prerequisitos, tanto en manejo de software como de conocimientos estadísticos. Han sido construidos con detalle, exponiendo cada uno de los conceptos y pasos utilizados. Los capítulos que les subsiguen, VI y VII, deben ser complementados con un aprendizaje estadístico especializado. No obstante, se establecen los elementos mínimos para ser comprendidos y replicados. Estos capítulos contienen menos detalles en los códigos, pues suponen la lectura de los capítulos anteriores.

En el capítulo VI se presenta un ejemplo de modelado de los ingresos. Para ello se utiliza una regresión lineal. Además, se evalúan los supuestos y se describen algunos de los problemas más comunes que se encuentran en este tipo de técnica y se reseñan métodos para resolverlos. De esta manera es posible dar cuenta de que los ingresos están diferenciados en el mercado de trabajo y que ello plantea desafíos metodológicos mucho más fuertes.

Finalmente, el capítulo VII se presenta en dos grandes secciones. La primera es una guía metodológica que nos muestra la utilización de la ENOE como un panel y, por tanto, expresa la lógica de los diferentes formatos que se pueden utilizar. Es una antesala para que, en la segunda sección, presentemos dos aplicaciones longitudinales con la base de datos: la creación de secuencias y las tasas de transición.

# **I. La Encuesta Nacional de Ocupación y Empleo (ENOE): sus temas y aplicaciones**

## **Introducción**

Para hacer un análisis estadístico —antes de incursionar en la parte práctica—, lo ideal es conocer muy bien la fuente de información: la Encuesta Nacional de Ocupación y Empleo (ENOE). En este tenor, en este primer capítulo se exponen los objetivos, historia y cambios de la ENOE desde su creación en el año 2005.

Del mismo modo, revisamos los grandes temas de la encuesta y su estructura en tanto fuente de información con unidades de información heterogéneas. Esto es importante porque un segundo objetivo de este capítulo es importar la información de la encuesta hacia el lenguaje de programación que utilizaremos. Por eso se establecen los comandos para importación y fusión o bien importar y fusionar para su uso transversal.

Para ello hemos dividido este capítulo en cinco secciones. En la primera hacemos una breve revisión de la historia de la ENOE. Posteriormente describimos sus elementos de diseño y básicos para comprender su estructura. En una tercera sección discutimos los componentes temáticos, y en la siguiente algunos cambios que deben tenerse en cuenta en caso de que se quiera utilizar esta fuente a lo largo del tiempo. Por último, reseñamos los temas y aplicaciones que se han llevado a cabo en los últimos años a partir de la revisión de textos publicados en contextos académicos.

## **A. Historia de la Encuesta Nacional de Ocupación y Empleo (ENOE)**

La Encuesta Nacional de Ocupación y Empleo es una encuesta trimestral que tiene por objetivo general “obtener información estadística sobre la fuerza de trabajo y las características ocupacionales

de la población a nivel nacional, estatal y por ciudades, así como de variables sociodemográficas que permitan profundizar en el análisis de los aspectos laborales” (INEGI, 2020, p. 9). Del mismo modo, se señala que esta fuente posibilita “ampliar la oferta de indicadores de carácter estratégico para el conocimiento cabal de la realidad nacional y la toma de decisiones orientadas a la formulación de políticas laborales” (INEGI, 2019, p. 9).

La ENOE tiene dos encuestas que le anteceden y mantienen una fuerte similitud. La primera es la Encuesta Nacional de Empleo (ENE), que se levantó de 1988 a 2000 de manera anual y posteriormente hasta 2004 de manera continua y trimestral. Por otro lado, la Encuesta Nacional de Empleo Urbano (ENEU), una encuesta semestral y trimestral, levantada entre 1987 a 2004, que ya incorporaba un diseño rotativo. De esta encuesta se heredó la posibilidad de poder hablar de ciudades autorrepresentadas.

No obstante, estas fuentes no son el único antecedente que tendrá México, un país pionero de la información. Las encuestas en México sobre la fuerza laboral datan de 1972, con la Encuesta Nacional de los Hogares (ENH). Posteriormente, de 1973 a 1974, se levantó la Encuesta Continua de Mano de Obra (ECMO); y de 1974 a 1984 la Encuesta Continua sobre Ocupación (ECSO).

La ENOE, como el resto de las encuestas mexicanas de empleo, mantiene como eje rector los marcos establecidos por la Organización Internacional del Trabajo (OIT), pero también incorpora elementos conceptuales recomendados por otros organismos como la Organización para la Cooperación y el Desarrollo Económico (OCDE); el grupo de París de Naciones Unidas (que trabaja empleo y remuneraciones); el grupo de Delhi de Naciones Unidas (para identificación de la informalidad) y la oficina de estadística de Naciones Unidas para los elementos sociodemográficos. Del mismo modo, también se siguen lineamientos de acuerdo con los Sistemas de Contabilidad Nacional (ISWGNA) y el Acuerdo Laboral de América de Norte (INEGI, 2019, p. 10).

La encuesta está diseñada para dar resultados de cada trimestre del año a nivel nacional, por entidad federativa y en 36 ciudades autorrepresentadas;<sup>1</sup> además, según la densidad de población, en lo-

<sup>1</sup> La encuesta inició con 32 ciudades de 2005 al primer trimestre de 2017, y continuó con 33 ciudades del segundo trimestre de 2017 al cuarto trimestre de 2018. Posteriormente, se

calidades de 100 000 y más habitantes, 15 000 a 99 999 habitantes, 2 500 a 14 999 habitantes, y en localidades de menos de 2 500 habitantes. No obstante, no todos los indicadores mantienen el mismo nivel de confiabilidad, por lo cual siempre se sugiere consultar los tabulados de precisión estadística que publica el INEGI con cada edición de la encuesta.<sup>2</sup>

## **B. Diseño y elementos básicos**

El diseño muestral de la Encuesta Nacional de Ocupación y Empleo se caracteriza por ser “probabilístico; por lo cual los resultados obtenidos de la encuesta se generalizan a toda la población, a su vez es bietápico, estratificado y por conglomerados, donde la unidad última de observación es la persona que al momento de la entrevista tenga 15 años cumplidos” (INEGI, 2019, p. 40).

La ENOE está compuesta por un panel rotativo. Cada individuo es entrevistado en cinco ocasiones y luego se va sustituyendo. Así, la “quinta parte de la muestra corresponde a viviendas que se visitarán por primera vez (llamado panel entrante) y el resto a las que van por la segunda a quinta visitas. Al realizarse esta última, sale de muestra ese 20% de viviendas (panel saliente) que son sustituidas por una cantidad igual (panel entrante)” (INEGI, 2019, p. 77). Sobre la construcción y lógica del panel se ahonda en el capítulo VII.

Se aplican dos cuestionarios a las personas: un Cuestionario Sociodemográfico (CS), que es el primero en aplicarse al iniciar la entrevista y cuya función es identificar a los integrantes del hogar, su sexo, edad, nivel de estudios y otros aspectos básicos de las personas, independientemente de que hagan o no una actividad económica. La batería de preguntas se mantiene igual desde 2005. De ahí que esta fuente sea ampliamente utilizada, para otros estudios que no necesariamente versan sobre mercados de trabajo.

---

hicieron estimaciones de 36 ciudades del primer trimestre al cuarto trimestre de 2019. Recientemente, se amplió a 39 ciudades en el primer trimestre de 2020.

<sup>2</sup> Revisar en el área de “Tabulados”, <<https://www.inegi.org.mx/programas/enoe/15ymas/default.html#Tabulados>>.

Y por otra parte, si bien la población en edad de trabajar (PET) es legalmente de 15 años y más, un segundo cuestionario, el de Ocupación y Empleo (COE), se aplica a las personas de 12 años y más, con preguntas que versan sobre la temática de la encuesta. Preguntar desde los 12 años permite continuidad histórica con otros ejercicios cuando la PET se definía más joven.<sup>3</sup>

Existen dos tipos de COE, uno básico y uno ampliado. El cuestionario ampliado se encuentra integrado por 11 baterías de preguntas, cada una con un objetivo y temática particular, mientras que en su versión básica cuenta con 9 baterías de preguntas. El ampliado se aplicó en todos los trimestres, desde el primer trimestre de 2005 hasta el segundo trimestre de 2006. Luego se empezó a aplicar el básico, a excepción del segundo trimestre de 2007 y de 2008, donde se aplicó el ampliado. Desde 2009, se aplica el ampliado en el primer trimestre de cada año (INEGI, 2017, p. 4). Es importante señalar que, por el carácter rotativo de la encuesta, sabemos que al menos una vez el individuo contestó la versión ampliada del cuestionario.

De estos dos cuestionarios (cs y COE) se obtienen cuatro tablas de información que operan con diferentes unidades de análisis, tal como se puede observar en la Figura 1-1. Hay una tabla de información llamada VIV que se refiere a las características que pueden ser recolectadas a nivel físico de la vivienda. Por ejemplo, la localización de ésta y otros elementos de acceso e infraestructura.

En la *tabla HOG* hay elementos que también pueden ser registrados a nivel de hogar. El hogar es una estructura social y dentro de una unidad física de vivienda es posible identificar más de un hogar. Se define hogar como “una organización estructurada a partir de lazos sociales entre personas unidas o no por relaciones de parentesco que comparten una misma vivienda (una sola persona también puede formar un hogar)” (INEGI, 2019, p. 21). De ahí que un grupo de personas se consideren un hogar si cumplen con la corredoría y un presupuesto común para la alimentación.

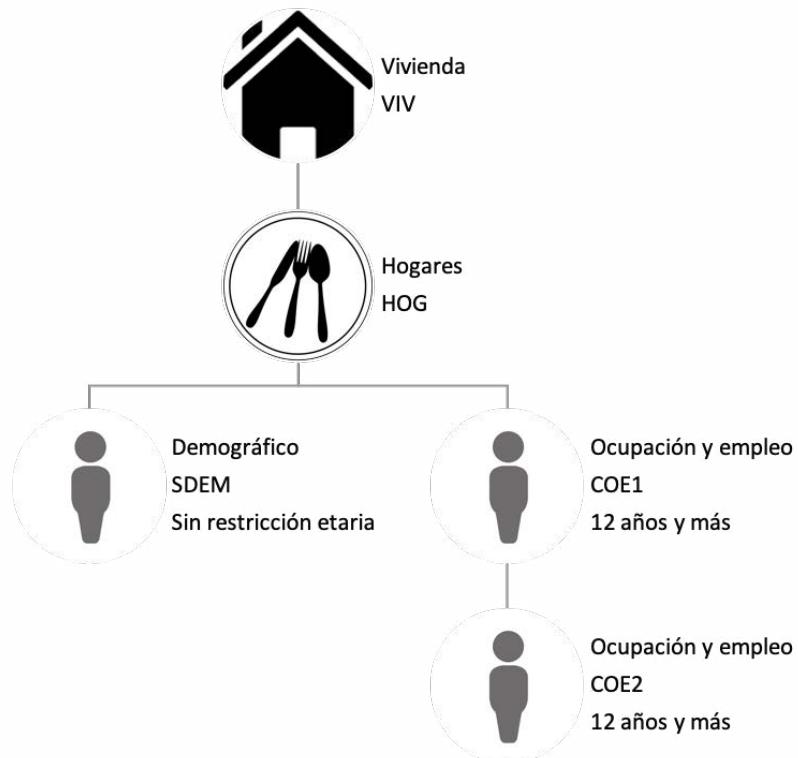
Los hogares están constituidos por una o más personas. Por eso, la unidad más pequeña de análisis es el individuo. Existe un grupo

---

<sup>3</sup> Con la modificación en 2015 a la Constitución Política de los Estados Unidos Mexicanos se estableció esta edad legal para trabajar.

de variables que tenemos para todas las personas ubicadas en la tabla SDEM, mientras que únicamente para las personas de 12 años y más tendremos una batería especial de preguntas que se organizan en las tablas COE1 y COE2.

**Figura 1-1. Esquema de las unidades de análisis en la ENOE y las tablas de información**



FUENTE: elaboración propia.

## C. Cobertura temática

Una vez que tenemos clara la estructura, es importante señalar qué variables e información se encuentran dentro de las tablas de la base de datos. Esta sección y la siguiente se basan en tres fuentes funda-

mentales: los microdatos, el sitio web (INEGI, 2020) y tres documentos metodológicos (INEGI, 2015, 2017 y 2019).

Para la población en general, se recolectan las características sociodemográficas que siguen: la relación de parentesco con el jefe del hogar; el sexo (mujer o hombre); la edad (en años cumplidos); la fecha de nacimiento y el lugar de nacimiento. También, para gente que supera la edad mínima escolar, se pregunta sobre el alfabetismo; el nivel de instrucción; carrera; antecedente escolar; egreso y asistencia escolar. Para personas de más de 12 años, también se pregunta la situación conyugal. Y para las mujeres de 12 años y más se pide información sobre el número de hijos nacidos.

Para la población en edad de trabajar (PET), la encuesta nos permite estudiar su condición de actividad económica (población económicamente activa, PEA, y la población no económicamente activa, PNEA). Del mismo modo, dentro de la PEA podemos estudiar la condición de ocupación (población ocupada y población desocupada). Dentro de la PNEA podemos identificar su condición de disponibilidad para trabajar (PNEA disponible y PNEA no disponible) y el tipo de actividades no económicas que realizan (si son estudiantes, se dedican a quehaceres domésticos, o son pensionados y/o jubilados, etcétera). Sobre esta estructura, en el capítulo III, en la Figura III-2, establecemos las relaciones entre tres variables que construye el INEGI.

En el caso de la población desocupada y de la PNEA, también podemos obtener información de la experiencia laboral de los desocupados.

Sobre las condiciones de la inserción laboral, la fuente es amplia. Se puede obtener información respecto a la ocupación principal, así como de la ocupación secundaria, en el caso de que la persona trabajadora tenga más de un trabajo.

De la *ocupación principal* podemos obtener el tipo de ocupación en términos de las tareas que desempeña; la posición en la ocupación de acuerdo con su condición de subordinación o no; el número de trabajadores en establecimiento del trabajador; el sector de actividad económica; si realiza sus actividades en el sector multinacional; prácticas contables; disponibilidad de local y lugar del trabajo. Del mismo modo, la encuesta sigue lineamientos internacionales y

el INEGI precodifica variables sobre el empleo informal y el sector informal (INEGI, 2014; Negrete, 2011).

Sobre la *jornada laboral* se puede obtener información de los días y horas trabajadas en la semana pasada, así como lo que habitualmente se trabaja a la semana. En caso de reportar no haber trabajado las mismas horas habituales en la semana pasada, se indaga sobre las causas. Del mismo modo, para conocer mejor el empleo temporal, se pregunta por los meses trabajados en el año. Además, hay una variable precodificada de las horas trabajadas.

Sobre las *remuneraciones* se tiene información de la forma de pago, el periodo de pago, así como el monto de ingresos mensuales. El INEGI además establece una variable precodificada sobre los ingresos laborales por hora y los ingresos mensuales.

Para la población ocupada también se tiene información sobre el acceso a atención médica por parte del trabajo. Y, además, se indaga acerca de la búsqueda de otro trabajo distinto al actual. En el cuestionario ampliado se incluyen variables de gran importancia para el estudio de la inserción laboral y sus condiciones, como la antigüedad en el trabajo actual; se brinda información del tipo de jornada de trabajo, el motivo por el que no trabaja todos los meses del año y la forma de conseguir el empleo.

En el caso de la población ocupada pero que lo hace de manera subordinada, es decir, que puede identificar a un jefe o identidad patronal, también se debe considerar otro tipo de análisis que se permite con esta fuente, como la condición de sindicalización; la condición de contrato de trabajo escrito; prestaciones laborales (i.e. vacaciones, aguinaldo, etcétera) cuya batería puede ser aún más amplia en el cuestionario ampliado.

La encuesta también brinda detalles para conocer mejor a la población desocupada. Nos permite estudiar la duración de la desocupación a partir de la fecha de inicio y término de la búsqueda de trabajo; la forma de búsqueda del trabajo y la experiencia laboral.

Para el resto de la PET, además de la batería sociodemográfica, se presenta información relevante sobre el uso del tiempo y, en el caso del cuestionario ampliado, se describe el acceso a apoyos económicos (becas, ayuda de programas y remesas). También hubo información disponible sobre lo que era el Seguro Popular.

Sin duda, la variedad temática de la fuente constituye un gran insumo para la investigación aplicada. No obstante, estos temas no siempre se han estudiado igual. La siguiente sección ahonda en estos cambios.

## D. Principales cambios en el tiempo

Existe una gran estabilidad en la encuesta en términos de su cobertura temática. Algunos reactivos han cambiado, pero los cambios más importantes están relacionados con algunos catálogos especializados que se han ido actualizando para tener compatibilidad con otras mediciones. A continuación mencionamos los cambios más importantes que han sido detectados en el uso de la base de datos y en la lectura de la documentación.

En el caso de los instrumentos de captación, durante casi quince años los cambios han sido pocos y en muchas ocasiones han sido adiciones, por lo que esta fuente es bastante comparable en el tiempo. Por ejemplo, en el cuestionario sociodemográfico las preguntas no han cambiado; pero en los cuestionarios básico y ampliado algunos cambios sí afectan la captación.

En 2007, un cambio fundamental fue la eliminación del filtro de la pregunta 2b: “¿En qué fecha fue la última vez que buscó trabajo?”, que limitaba la información a los que buscaban trabajo durante un mes antes de la entrevista, de ahí que con esta modificación haya más información sobre los buscadores. En el caso del ampliado en este año se incorporaron tres preguntas para caracterizar el trabajo secundario (preguntas 7e, 7f y 7g).

En 2013, se incluyeron nuevas actividades a la batería del uso del tiempo, se mantienen las anteriores y se agrega el traslado o acompañamiento, así como la realización de compras y trámites.

Los catálogos permiten identificar a través de códigos los valores de una cantidad de respuestas. Normalmente, los catálogos son utilizados para homogenizar información entre fuentes e incluso, en algunos casos, permite hacer comparaciones internacionales de manera más sencilla. Del mismo modo, los catálogos permiten movernos en niveles de granularidad. Algunos catálogos pueden trabajarse en diferentes niveles de dígitos, lo cual permite tener

grupos más o menos extendidos de acuerdo con el número de cifras en los códigos, por lo que hay que ser cuidadosos y cuidadosas al usar preguntas a lo largo del tiempo que utilicen catálogos diferentes. Puede ser que la pregunta no se cambie en su fraseo y que en el cuestionario aparezca idéntica, pero los valores de las respuestas pueden ser diferentes.

Dado que el mercado de trabajo y la organización cambia, estos catálogos necesitan actualizarse. En 2008 se realizó una actualización al catálogo de franquicias y en 2009 se actualizó el de Dependencias e Instituciones de Interés Público. En 2012 se hicieron también varios cambios en los catálogos: se dejó de usar la Clasificación Mexicana de Ocupaciones (CMO) para la pregunta de tareas y actividades en la ocupación (página 3 del cuestionario básico y ampliado) con el Sistema Nacional de Clasificación de Ocupaciones (SINCO). En el mismo año, se cambió el catálogo de carreras por el de la Clasificación Mexicana de Programas de Estudio por Campos de Formación Académica (CMPE), la cual se volvió a actualizar en 2016.

Asimismo, se han realizado actualizaciones en el catálogo de parentesco, así como en el de lugar de nacimiento a lo largo de estos años, por lo que recomendamos tomar precauciones con el uso longitudinal de estas variables.

En el caso de los microdatos, es decir, la información en la unidad más pequeña, ha habido un par de cambios. En 2013 se ajustaron los resultados a las proyecciones demográficas del CONAPO con base en los resultados del Censo de Población y Vivienda 2010.

Del mismo modo, el INEGI reporta que hizo algunos cambios en los datos en formato de STATA y de SPSS (pero esto no afecta lo publicado en otros formatos). A inicios de 2019, se reemplazó el archivo COE2 (Cuestionario de Ocupación y Empleo) de las bases de datos correspondientes a los trimestres primero, segundo y tercero de 2018, debido a un ajuste de la etiqueta del campo de la pregunta 6b2.

Si se toman en cuenta estos elementos metodológicos y temáticos, quien lee el texto puede hacerse una idea de las múltiples aplicaciones que tiene una encuesta de este tipo. En el siguiente acápite se reseñan algunos trabajos donde se ha utilizado esta encuesta, señalando sus temas y orígenes de aplicación.

## E. Reseña de los estudios que usan la Encuesta Nacional de Ocupación y Empleo en esta década<sup>4</sup>

México ha tenido una larga trayectoria académica sobre el estudio de su mercado de trabajo. Sin duda, hay trabajos seminales que han dado cuenta históricamente de los procesos de transformación que ha vivido el país y sus mercados de trabajo desde los años 1970 hasta inicios de este siglo (García, 2012; García & Pacheco, 2011; Pacheco, 2004; Pedrero, 2003 y 2018; Pedrero & Rendón, 1982; Rendón & Salas, 1986; Salas & Zepeda, 2003).

En la presente sección se reseñan los temas abordados, enfoques conceptuales y metodológicos de 160 trabajos publicados en inglés y español desde 2010.<sup>5</sup> Sin embargo, estos textos no son exhaustivos en lo que concierne a la investigación sobre la Encuesta Nacional de Ocupación y Empleo (ENOE), ya que probablemente existen otros productos de investigación tales como tesis, documentos de trabajo o conferencias en congresos, por ejemplo, que por carecer de registro son difíciles de rastrear. O bien, seguramente hay libros escritos en idiomas distintos al español e inglés, así como documentos no publicados de organismos internacionales. No obstante, la muestra que presentamos sí genera un cúmulo de voces disponible para la consulta y nos muestra la conversación que se está llevando a cabo en el ámbito académico respecto al empleo y la ocupación en México. En este sentido, es evidente la variedad de temas que se están estudiando en el marco de los mercados de trabajo e, incluso, se muestra cómo la encuesta puede ser utilizada para otros temas diferentes a la ocupación y el empleo.

La presente sección se organiza de la siguiente manera: en un primer apartado hablamos de números, y de cuántos trabajos y cuándo se realizaron. En un segundo apartado, sobre las herramientas de análisis de texto; mostramos, a partir de los resúmenes de los trabajos y las palabras clave registradas, los temas que resultaron más recurrentes en la última década. Finalmente, en el tercer apartado hacemos una revisión de los métodos y la perspectiva de los trabajos.

<sup>4</sup> Agradezco a Fernanda Morfín su colaboración para sistematizar la información utilizada en este capítulo.

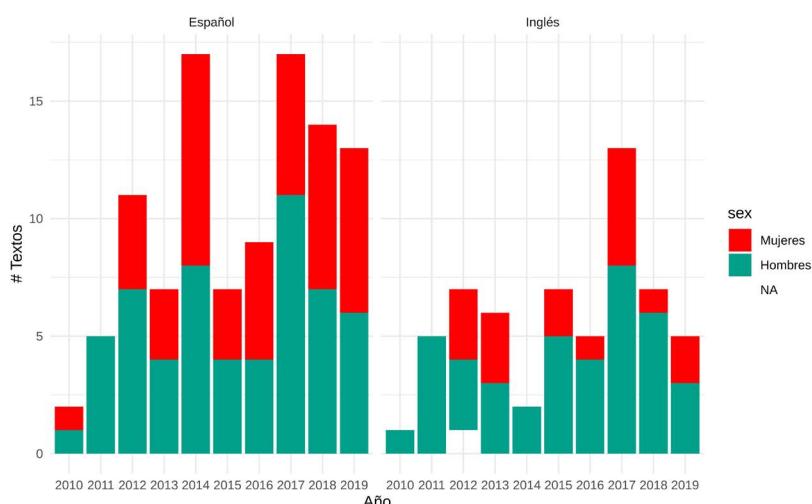
<sup>5</sup> La bibliografía de estos textos se encuentra en el anexo al final del libro, así como el código utilizado para el análisis de texto.

## a. Numeralia

Los textos analizados los encontramos en los buscadores de *Google Scholar*. Dimos prioridad a los que utilizan a la ENOE como fuente principal o fundamental en el análisis por lo que analizamos 160 textos publicados entre 2010 y 2019. La mayor parte de ellos fueron documentos escritos en español (102 textos, 63.75%) y un grupo de 58 textos (36.25%) en inglés.

En la mayoría de los casos, los textos analizados fueron escritos en comunidad. En promedio, cada uno tiene 1.94 autores, con un máximo de hasta cinco. Sin discriminar por posición de autor, las colaboraciones son mayormente de hombres (59.9%), mientras que las mujeres participan en 40.1%. Esta distribución se mantiene muy similar cuando nos quedamos con los primeros autores, siendo los hombres quienes participan en el 61% de los textos y las mujeres en un 39%.

**Gráfica I-1. Número de textos según idioma y sexo del primer autor (2010-2019)**



FUENTE: elaboración propia. En 2012, en inglés se incluye un texto de autoría institucional que no tiene sexo.

En la Gráfica I-1 se observa cómo se distribuyen los textos según idioma y sexo del primer autor. Por un lado se observa un mayor número de textos en español y, por otra parte, se puede apreciar que la autoría de las mujeres es menor a la de los hombres. La brecha por sexo se amplía para la literatura en inglés.<sup>6</sup>

Como se puede ver, no hay un patrón de tendencia en la publicación, aunque resalta una mayor cantidad de textos en español para 2014, y en inglés para 2018.

Una vez que describimos de manera general la muestra y sus autores, es imperativo movernos al análisis de los temas estudiados en estos textos.

### b. Los temas de la última década

En un primer análisis haremos una revisión de los temas reportados en las palabras clave y luego, desde una perspectiva más amplia, analizaremos los resúmenes de los textos.<sup>7</sup> Para dicho análisis utilizaremos las herramientas de lenguaje natural provistas por el paquete *UDPipe* (Straka & Straková, 2017).

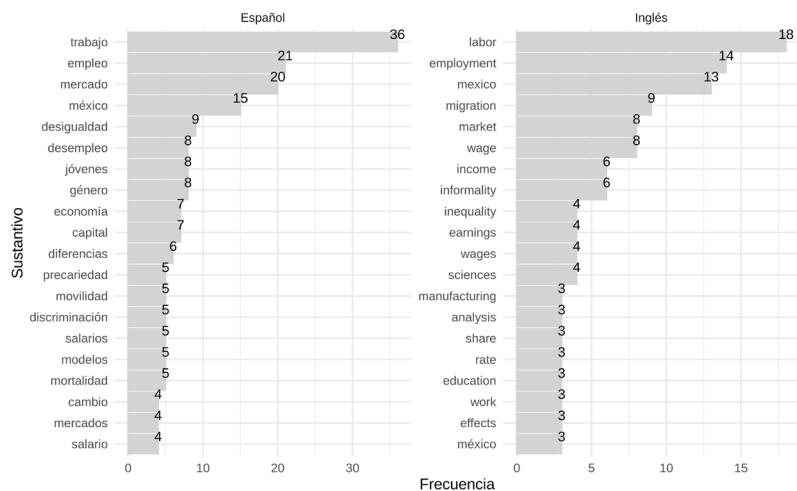
En el caso de las palabras clave escogidas por los autores, tenemos diferencias entre lo que se publica en inglés y en español en términos de los temas. Los sustantivos más utilizados en estas palabras clave se presentan en la Gráfica I-2.

En ambos idiomas, el trabajo y el empleo lideran los sustantivos más utilizados en las palabras clave. Del mismo modo, “mercado” es una palabra que sobresale. Es notable en la Gráfica I-2 cómo la desigualdad se presenta como un tema común en la literatura en español (número 5 en el *ranking* de sustantivos), pero también tiene su contraparte en la literatura en inglés (“inequality”, *ranking* 11). En la literatura en español, podemos observar que se presentan grupos de análisis como palabras clave, tales como género y jóvenes. Sobre fenómenos del mercado de trabajo, podemos observar que mientras en la literatura en español se mencionan en varias ocasiones el desempleo, la discriminación y la precariedad, en la literatura en

<sup>6</sup> Para la literatura en español, dentro del periodo en estudio, la participación de autoras (primer autor) es 44.1% mientras que en inglés esto disminuye a 29.31%.

<sup>7</sup> 33 de los textos revisados no incluían palabras clave reportadas por los autores.

**Gráfica I-2. Palabras (sustantivos) más utilizadas en las palabras clave (2010-2019)**



FUENTE: elaboración propia.

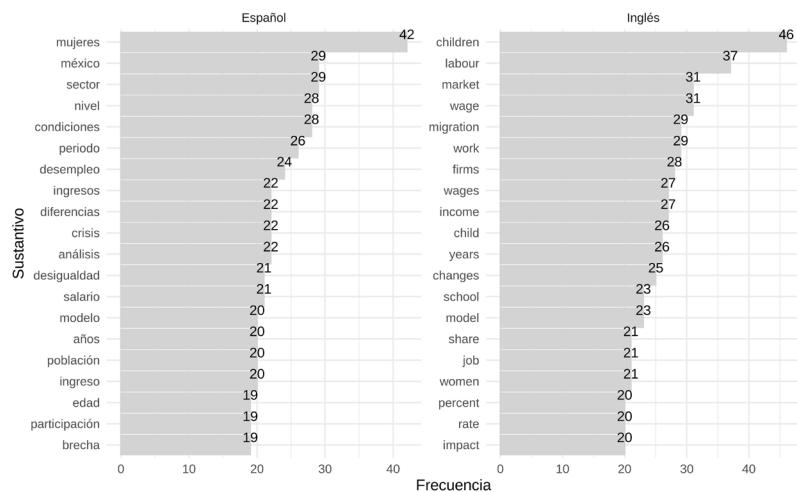
en inglés hay mucha más presencia de análisis sobre las remuneraciones (*earnings*, *wages*, *income*). De igual manera, la importancia de estudiar la migración con la ENOE como fuente de datos está más presente en la literatura en lengua sajona. Resalta en este análisis que en la literatura en inglés hay mucha mayor presencia del enfoque de la informalidad que en la literatura en español. Otro tema que parece ser importante es la vinculación de trabajo y educación: aparece en cuatro trabajos en español y tres en inglés.

Este primer análisis nos presenta hacia dónde se dirigen los productos académicos en términos de las problemáticas sociales. Para ampliar lo establecido e incorporar a todos los textos, aplicamos este análisis pero considerando en esta ocasión los resúmenes (Gráfica I-3).

Los temas y las diferencias por idioma en la producción académica son más evidentes en la Gráfica I-3 que en la Gráfica I-2. En esta nueva gráfica, resalta más el elemento del género en el mercado de trabajo, con una alta presencia de la palabra “mujer” y “género”. Del mismo modo, en la literatura en español resalta también la inclu-

sión de la palabra “crisis”, que se utiliza 22 veces en los resúmenes analizados, elemento que no había sido evidenciado con las palabras clave. Sigue presente la desigualdad como un tema recurrente, no sólo por la presencia de “desigualdad”, sino que también aparecen las palabras “diferencias” y “brechas” dentro de las 20 palabras (sustantivos) más utilizadas. Es posible observar que se estudian elementos de inserción laboral a través de la palabra “condiciones”, pero también se estudia la exclusión. “Desempleo” sigue siendo una palabra importante, al igual que la de “participación”, que da cuenta de la entrada a la PEA.

**Gráfica I-3. Sustantivos más utilizados en los resúmenes de los textos según idioma de publicación (2010-2019)**



FUENTE: elaboración propia.<sup>8</sup>

<sup>8</sup> Para la gráfica se excluyeron algunos sustantivos comunes pero que no dejaban evidenciar al resto. Para el caso de las publicaciones en español, se excluyeron: “trabajo”, “empleo”, “objetivo”, “documento”, “méjico”, “mexico”, “resultados”, “trabajadores”, “ocupación”, “presente”, “artículo”, “estudio”, “parte”, “partir”, “datos”, “encuesta”, “ENOE”, “mercado”. Para el caso de las publicaciones en inglés se excluyeron: “employment”, “workers”, “mexico”, “survey”, “data”, “paper”, “document”, “results”, “labor”, “sector”.

En el caso de las palabras más utilizadas en inglés, al analizar todos los resúmenes y no sólo las palabras clave, se observa más la preocupación por el vínculo educación y trabajo con “school”; del mismo modo, se evidencia la importancia de grupos dentro del mercado de trabajo de manera más clara como las mujeres (“women”) y los niños y niñas (“child”).

Para mostrar un poco más cómo se manejan los temas, además del análisis de los sustantivos, presentamos las palabras clave que extrae el algoritmo RAKE (acrónimo de Rapid Automatic Keyword Extraction). Este algoritmo busca palabras clave buscando una secuencia contigua de palabras que no contengan palabras irrelevantes.

Es decir, al calcular una puntuación para cada palabra que forma parte de cualquier palabra clave candidata el algoritmo observa cuántas veces aparece cada palabra y cuántas veces se produce con otras palabras. Cada palabra obtiene un puntaje que es la razón del grado de la palabra (cuántas veces se produce con otras palabras) a la frecuencia de la palabra. Se calcula un puntaje RAKE para la palabra clave candidata completa, sumando los puntajes de cada una de las palabras que definen la palabra clave. Los resultados se pueden ver en el Cuadro I-1, donde se presentan sólo las palabras que implican un enlace con más de una palabra (Straka & Straková, 2017).

**Cuadro I-1. Detección de palabras clave según el algoritmo RAKE, las 20 más frecuentes según idioma**

	Palabras clave/ Español	Frecuencia	Puntaje RAKE	Palabras clave/ Inglés	Frecuencia	Puntaje RAKE
1	encuesta nacional	50	2.01	child labour	14	2.40
2	mercado laboral	21	1.90	informal sector	13	2.37
3	sector informal	13	2.09	national survey	10	2.59
4	desigualdad salarial	11	1.83	formal sector	10	2.39
5	capital humano	10	2.29	educated workers	10	1.71
6	condiciones laborales	10	1.92	informal employment	10	1.69
7	brecha salarial	9	2.06	health insurance	7	3.11

	Palabras clave/ Español	Frecuencia	Puntaje RAKE	Palabras clave/ Inglés	Frecuencia	Puntaje RAKE
8	instituto nacional	8	2.04	informal workers	7	1.76
9	primer trimestre	8	1.85	mexican labor market	6	4.62
10	frontera norte	7	1.69	labor market	6	3.11
11	cuenta propia	7	1.53	child labor	6	2.58
12	nivel educativo	6	2.05	unemployment rate	6	2.43
13	precariedad laboral	6	1.71	labour share	6	2.10
14	diferencias salariales	6	1.64	mexican economy	6	2.01
15	salario mínimo	6	1.49	children aged	6	1.34
16	industria manufacturera	5	2.30	labor markets	5	2.79
17	jornada laboral	5	1.90	formal jobs	5	2.03
18	principales resultados	5	1.08	skilled workers	5	1.57
19	mercado laboral mexicano	4	3.40	school attendance	5	1.40
20	vulnerabilidad laboral	4	2.07	informal sector jobs	4	3.27

FUENTE: elaboración propia.

Este análisis complementa lo que ya habíamos analizado y observamos que destacan algunos elementos sectoriales. Por ejemplo, en el caso de la literatura en español podemos observar que existe una fuerte atención a la frontera norte y a la industria manufacturera. De nuevo podemos encontrar evidencia de cómo la desigualdad es un tema central en los resúmenes: “diferencias salariales”, “brecha salarial” y “desigualdad salarial”. En este análisis observamos cómo la informalidad sigue siendo estudiada desde la literatura en español, lo cual no se evidenciaba en los análisis anteriores. Pero además del esquema de la desigualdad, podemos observar que otros esquemas analíticos, tales como la precariedad laboral y la vulnerabilidad, también están presentes en el análisis del mercado de trabajo mexicano.

En el caso de la extracción para lengua inglesa, tenemos el tema de la calificación de los trabajadores (“skilled workers”), elemento que no había aparecido en los anteriores análisis, y el estudio del seguro de salud (“health insurance”).

Este pequeño análisis da cuenta que, en la última década, esquemas como la informalidad, siguen vigentes en los análisis del mercado de trabajo; mientras que otros también avanzan, como es el caso de la precariedad y el de la vulnerabilidad. Pareciera que el tema al que más frecuentemente se hace referencia es la desigualdad. Al mismo tiempo, los estudios se realizan para grupos específicos, en los que resaltan los de mujeres, jóvenes, niños y migrantes.

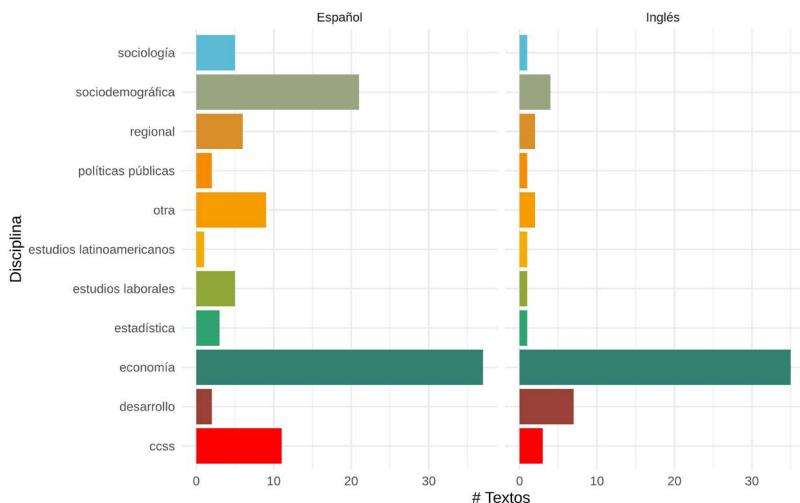
Estos temas también se enmarcan en diferentes disciplinas. A continuación revisaremos desde qué disciplina y métodos se llevan a cabo los análisis.

### **c. Desde dónde y cómo se estudia**

En la subsección anterior revisamos los temas, ahora veremos cómo estos temas se posicionan desde diferentes espacios disciplinarios. Tomando en cuenta la naturaleza de la publicación del artículo, se codificó la disciplina; es decir que lo analizado refiere a la revista, editorial u organismo que publica la investigación.

Para fines de exposición se han hecho algunas agrupaciones en términos de la definición que las revistas o entes publicadores hacen sobre ellos mismos. En las ciencias sociales se mantienen las publicaciones donde la revista tiene un nombre general que incluye “Ciencias sociales”, “Estudios sociales” o “Ciencias sociales y humanidades”. De ningún modo esta separación tiene que ver con que la sociodemografía o la economía no sean ciencias sociales.

Resalta que la mayor parte de los textos que usan la encuesta lo hacen desde una perspectiva económica (45.0%), seguida por la sociodemográfica (15.6%) y las publicaciones en el marco de las ciencias sociales (8.7%). No obstante, existen trabajos escritos desde la computación, la actuaria, las políticas públicas, la medicina y la salud que han utilizado la fuente de información. Ello da cuenta de la variedad de aplicaciones que tiene la encuesta.

**Gráfica I-4. Número de textos según disciplina e idioma**

FUENTE: elaboración propia. En “otra” se incluyen textos de medicina, salud, computación, migración, estudios rurales y multidisciplinarios.

Es claro que la distribución de los textos analizados tiene una diferencia por disciplina y por idioma, lo que puede explicar las diferencias temáticas analizadas en los apartados anteriores. Resalta que el universo analizado en español es más variado en términos disciplinares. De los textos analizados en español, un 36.3% están publicados en un espacio económico, frente a un 60.3% de los textos analizados en inglés. En español es aún más importante la presencia de los estudios desde la perspectiva sociodemográfica y de ciencias sociales que en inglés, mientras que la perspectiva de desarrollo es más importante en la literatura en inglés.

El elemento disciplinar marca los métodos. A continuación, también con un análisis de extracción de palabras clave con el algoritmo RAKE, presentamos algunos resultados de los métodos más utilizados. Es importante hacer notar que cada investigación puede incorporar más de un método. De esta manera obtenemos el Cuadro I-2.

**Cuadro I-2. Detección de palabras clave según algoritmo RAKE, las 10 más frecuentes**

Palabras clave	Frecuencia	Puntaje RAKE
análisis descriptivo	27	1.90
panel	15	0
regresión logística	14	2.15
modelo	11	0.55
descomposición	7	0
regresión lineal	6	2.40
series	5	0
tiempo	5	0
datos	4	0
diferencia	4	0
regresión cuantílica	3	2.03
efectos fijos	3	2
variables instrumentales	3	2
equilibrio general	3	1.75

FUENTE: elaboración propia.

El análisis descriptivo sigue jugando un papel muy importante en las investigaciones, colocándose como el más común. El análisis de panel o construcción de un panel también es relevante. Aunque se debe hacer la salvedad de que algunas de estas investigaciones que utilizan paneles, muchas veces refieren a espacios geográficos y grupos, no al panel mismo de la encuesta a nivel individual.

La palabra “modelo” se presenta sola en el Cuadro I-2 debido a los diferentes modelos que no lograron crear una conexión (11 veces) con ningún otro método, y de ahí que el algoritmo dejara la palabra sola. No obstante, esto implica que, pese a que el análisis descriptivo sigue siendo importante, también hay una alta tendencia a producir modelos. Dentro de los modelos específicos que se utilizan están la regresión logística y lineal. Se debe recordar que se presentaba “mínimos cuadrados ordinarios” en referencia a la regresión lineal, incluida también en el Cuadro I-2. Dado que uno

de los temas más estudiados es la desigualdad, hay varios trabajos que tienen elementos de descomposición, ya sea de brechas o de índices de desigualdad.

Esta pequeña revisión nos permite dar paso al resto de capítulos de este libro, pero instamos además que se consulte el cúmulo de bibliografía que recolectamos pues es ya, de por sí, un avance en el estudio de los mercados de trabajo en México y se puede consultar en el anexo de este libro. A lo largo de algunos ejemplos aplicados, se irá retomando parte de los trabajos para quien desee profundizar sobre estos temas y técnicas analíticas.

El siguiente capítulo presenta la fuente de datos. La introducción al análisis descriptivo se ve en los capítulos III y IV. Mientras que en el resto se exponen la inferencia, los modelos de regresión lineal e introducen al panel propio de la encuesta. Sin duda, estos elementos coinciden con las estrategias más usadas en el estudio de los mercados laborales.

## **II. Una muy breve introducción a R y preparación de la fuente de información**

### **Introducción**

El presente libro utiliza el *software R* y la plataforma *RStudio*, por lo cual se necesitan instalar ambos programas. Este texto fue preparado con la versión 4.0.2 (2020-06-22) *Taking off Again* de *R* y con la versión 1.3.1056 de *RStudio*. *R* puede ser descargado de uno de los sitios “espejo” (*mirror sites*) en <<http://cran.r-project.org/mirrors.html>>. Para una mejor instalación se sugiere escoger un espejo cercano a la locación. Si bien la revisión de técnicas estadísticas puede hacerse con cualquier otro paquete estadístico, se eligió éste por ser de libre acceso y conocimiento colectivo.

En cuanto a *RStudio*, puede descargarse de <<https://rstudio.com/products/rstudio/download/>>. Es conveniente utilizar la primera opción con licencia *open source* ya que no se destinará a uso comercial.

*RStudio* es un IDE (Integrated Development Environment) bastante popular y compatible con *tidyverse*, un grupo de paquetes en expansión dentro de la comunidad de usuarios de *R*. En la medida de lo posible intentamos usar un tipo de programación compatible.

Para instalar ambos programas se debe contar con un equipo con al menos 2 GB de RAM, dado que los códigos que se muestran están diseñados para trabajar con la base completa de la ENOE. Esto se vuelve aún más necesario en el último capítulo donde trabajamos con las cinco ediciones de la encuesta para utilizar el panel rotativo. Del mismo modo, se sugiere tener al menos 1 GB de memoria liberada para el almacenamiento de la encuesta y los subproductos que se puedan ir creando a lo largo del procesamiento.

En cuanto a los sistemas operativos que soporta *R*, se recomienda Windows 7 o superior y, de preferencia, de 64 bits. En el caso de Mac OS X, una versión 10.13 (High Sierra) o superior. Para quienes usan Linux, se sugiere una versión de Ubuntu 16.04 o superior y una

versión de CentOS / Red Hat Enterprise Linux 7.x, 8.x o mayor, de preferencia con un procesador de 64 bits.

Dentro del funcionamiento de estos códigos, se ha trabajado además con una serie de paquetes que deben instalarse. A continuación, reseñamos los más importantes para que el usuario se vaya familiarizando con ellos.

## A. Paquetes utilizados

*R* funciona en dos partes “base”: lo necesario para correr *R* y algunas funciones básicas, con herramientas preinstaladas. Además, hay más de 17 000 paquetes en CRAN (Comprehensive R Archive Network), los cuales se pueden encontrar en <<https://cran.r-project.org/web/packages/index.html>>.

En este libro utilizaremos el conjunto de paquetes *tidyverse* (Wickham, Averick, Bryan, Chang, McGowan *et al.*, 2019) por tener una amplia relación con la interfase *RStudio*, además de ser parte de un movimiento que plantea nuevas formas de programar. En general, para programar vamos a apoyarnos del formato *tidy*, utilizando muchos comandos del paquete *dplyr*.

Según sus creadores, el objetivo del *tidyverse*:

[...] es proporcionar herramientas para los desafíos más comunes; no para resolver todos los problemas posibles. Notablemente, el *tidyverse* no incluye herramientas para el modelado estadístico o la comunicación. Estos kits de herramientas son críticos para la ciencia de datos, pero son tan grandes que merecen un tratamiento por separado. El paquete *tidyverse* permite a los usuarios instalar todos los paquetes *tidyverse* con un solo comando (Wickham, Averick, Bryan *et al.*, 2019).

Además de este gran grupo de paquetes, utilizaremos otros complementarios:

- El paquete *broom* (Robinson & Hayes, 2019), de amplia utilidad para almacenar la información de resultados estadísticos en formatos compatibles con *tibbles*.<sup>9</sup>

---

<sup>9</sup> Son tablas en formato *tidy*, utilizado en *tidyverse*.

- El paquete *car* (*Companion to Applied Regression*) (Fox, Weisberg, Price, Adler, Bates *et al.*, 2019) con funciones que facilitan la aplicación e interpretación del análisis de regresión —que manejaremos en el capítulo VI, además de las paquetes *lm.beta* (Behrendt, 2014) y *robustbase* (Maechler, Rousseeuw, Croux, Todorov, Ruckstuhl *et al.*, 2020).
- El paquete *DescTools: Tools for Descriptive Statistics* (SIGNORELL, AHO, ALFONS, ANDEREgg, ARAGON *et al.* 2020), caja de herramientas de estadísticas descriptivas de la cuales retomamos un par de funciones.
- Para algunas extensiones de manejo de bases de datos, también se utiliza el paquete *extdplyr* (Wang, 2017).
- También se usa como alternativa para crear gráficas el paquete *esquisse* (Meyer, Perrier & Caroll, 2020).
- En general, importamos las bases de datos de la ENOE desde su formato de Stata para aprovechar el etiquetado de la institución que las publica. Para ello, usamos el paquete *haven* (Wickham & Miller, 2019).
- En el capítulo IV revisamos algunas medidas de desigualdad para estudiar los ingresos laborales. Para ello usamos el paquete *ineq* (Zeileis & Kleiber, 2014); y para las gráficas el paquete *gglorenz* (Chen, 2020).
- Para un mejor manejo de tabulados y porcentajes, utilizamos el paquete *janitor* (Firke, Denney, Haid, Knight & Grosser, 2020). Asimismo, este paquete tiene funcionalidades de limpieza de nombres de columnas.
- Al inicio de cada capítulo práctico, utilizamos el comando “*p\_load()*” de la paquetería *pacman* (Rinker, Kurkiewicz, Hughitt, Wang, Aden-Buie, Wang & Burk, 2019). Esto nos asegura que se instalen y carguen todos los paquetes necesarios.
- Para las escalas de color, utilizamos dos paquetes: *RColorBrewer* (Neuwirth, 2014) y *wesanderson* (Ram, Wickham, Richards & Baggett, 2018).
- El manejo de etiquetas y su edición se hace con el paquete *sjlabelled* (Lüdecke & Ranzolin, 2020).
- Para producir gráficas de estimaciones de modelos estadísticos se utiliza el paquete *sjPlot* (Lüdecke, 2020).

- Para las tablas de estas estimaciones también empleamos la paquetería *stargazer* (Hlavac, 2018b).
- Para los factores de expansión y el diseño muestral se utilizó el paquete *srvyr* (Ellis, Lumley, Žóltak & Schneider, 2020).
- El capítulo VII, que versa sobre análisis de secuencias, se realiza con los aportes de *TrAMineR* (Gabadinho *et al.*, 2020) y las gráficas con *easyalluvial* (Koneswarakantha, 2020).

Finalmente, la integración del código a formato de documento para la escritura de este libro se hizo con la ayuda del paquete *Rmarkdown* (Allaire, Xie, McPherson, Luraschi, Ushey *et al.*, 2020; Xie, Allaire & Grolemund, 2018).

## B. Elementos básicos

En esta sección revisaremos los elementos básicos necesarios para el manejo de las bases de microdatos de empleo. Hablaremos un poco de la instalación, y cómo funcionan los comandos más simples.

En *RStudio* podemos tener varias ventanas que permiten un mayor control de nuestro “ambiente”, el historial, los *scripts* o códigos que escribimos y, por supuesto, la consola que también tiene el símbolo “>” con *R*.

```
2+2
## [1] 4
```

En los recuadros grises, podemos ver el código y el resultado. Como se observa, después de “##” se indica lo que se obtiene en la consola del paquete como resultado. De esta manera, quien lee puede observar el resultado de las operaciones a lo largo del libro.

Sin duda, aprender *R* es mucho más sencillo de la mano de los productores de paquetes que más utilizamos, por ello recomiendo a cualquiera que se inicie en este paquete estadístico la versión en español de “*R* para ciencia de datos”, disponible en <<https://es.r4ds.hadley.nz/>>.

A continuación, listamos algunos elementos básicos para introducir el presente libro de una manera más adecuada.

## a. La paquetería de *R*

Como señalamos al inicio de este capítulo, algunas funciones necesarias para nuestro análisis deben ser instaladas. Los paquetes publicados en el repositorio CRAN están avalados por una comunidad y han sido revisados. Estos se pueden instalar directamente desde la consola o desde nuestro *script*.

Por ejemplo, podemos instalar el paquete *foreign* con el comando “`install.packages()`”. Dentro de cada función de *R*, podemos colocar argumentos separados por comas. En este caso colocamos además un argumento que es muy útil, “`dependencies=TRUE`”, que nos instala cualquier otra paquetería necesaria para *foreign*, que normalmente llamamos dependencias.

```
install.packages("foreign", dependencies = TRUE) # Instala un paquete específico
```

Al instalar paquetes, es normal que haya letras de otro color en la consola; esto quiere decir que *R* está avisando de los cambios que hace en el equipo y no es un error, a excepción del mensaje “`installation of package had non-zero exit status`”, que indica que se ha omitido el resultado en la consola porque puede ser muy largo.

```
library(foreign) # Cargar La Librería
```

Instalar un paquete es una acción de una sola vez. Pero *R* tiene una característica llamada en inglés *lazy loading*, ello quiere decir que carga de manera perezosa, o lo que algunos llaman “carga diferida”. Por lo que, a pesar de haber instalado un paquete, hay que “llamarlo” en cada sesión de nuestro programa para utilizar los comandos que provienen de él. Esto parece molesto, pero en realidad es muy eficiente para el uso de memoria. Por ello es una buena sugerencia cargar los paquetes que se vayan a utilizar. Cuando aplicamos este comando no tenemos ningún resultado en la consola.

Los paquetes, al estar “vivos” dentro de la comunidad de usuarios y desarrolladores, son sujetos a modificaciones, por lo que es conveniente actualizarlos de vez en cuando. Puedes hacerlo para paquetes específicos, colocando el nombre en comillas del paquete

dentro de la función, o bien, si quieras hacerlo para todos, puedes hacerlo de la siguiente manera:

```
update.packages()# Actualizar librerías instalados
```

El paquete *pacman*, con su comando “*p\_load()*”, nos permite cargar paquetes y si no los tenemos, los instala. Puede ser útil.

Si quieras instalar todos los paquetes usados a lo largo del libro, puedes ejecutar los siguientes códigos:

```
if (!require("pacman")) install.packages("pacman") # instala pacman si se requiere
## Loading required package: pacman

pacman::p_load(tidyverse, # conjunto de paquetes tidy
               broom, # paquete para adecuar resultados estadísticos
               car, # para la regresión lineal
               DescTools, # caja de herramientas estadísticas
               esquisse, # para hacer ggplot con drag and drop
               extdplyr, # extensión de dplyr
               gglörenz, # gráfico de curva de Lorenz
               haven, # importa archivos desde formatos .dta y .sav
               ineq, # Para medidas de inequidad y desigualdad
               janitor,# para tabulado y limpieza de nombres
               lm.beta, # para coeficientes beta
               Rcolorbrewer, #paletas de color
               wesanderson, #paletas de color películas Wes Anderson
               robustbase, # Para estimaciones de modelos robustos
               sjlabelled, #manejo de etiquetas y edición
               sjplot, #graficos de estimaciones de modelos
               stargazer, #Tablas de estimaciones de modelos
               srvyr, # diseño muestral
               TraMiner # Para el análisis de secuencias
               )
```

En el capítulo VII utilizamos un paquete que no está aún publicado en CRAN, para hacer gráficas aluviales; es un paquete en desarrollo. Para poder instalar este tipo de paquetes, se hace con ayuda de la paquetería remotes, y luego se puede instalar desde el repositorio de GitHub donde esté alojado el paquete:

```
if (!require("remotes")) install.packages("remotes") # instala remotes si se requiere
## Loading required package: remotes

remotes::install_github("erblast/easyalluvial")
```

Cuando instalamos de esta forma, es posible que se pregunte por actualizar algún paquete, a lo cual es recomendable decir que sí, con

“Y”, de “Yes” en inglés; así como si deseas instalar algún paquete que se compile, también hay que poner que sí, es decir, una “Y”.

### b. ¿Dónde estamos trabajando?

Para tener a la mano nuestros archivos y demás material de trabajo, podemos establecer el directorio de trabajo. Primero habrá que verificar en nuestra computadora la ruta donde estamos trabajando:

```
getwd() # Directorio actual  
## [1] "/Users/Libro/Código"
```

Para establecer la ruta se utiliza el comando “setwd()”. Esta es una operación que se ordena fácilmente desde el menú de *RStudio*: “Session → Set Working Directory → Choose Directory” → carpeta de elección.

```
setwd("/Users/Libro/Código") # cambia La ruta al directorio que se escribe
```

Otra forma de trabajar es creando proyectos. Para crear proyectos cuyas rutas sean relativas, es decir, que no dependan de la ruta global de la computadora sino de la carpeta donde está el proyecto, recomendamos revisar la sección 8 del texto “R para ciencia de datos”, disponible en <<https://es.r4ds.hadley.nz/flujo-de-trabajo-proyectos.html>>.

### c. Comandos básicos para empezar

A continuación, dejo una lista de operaciones simples que se pueden ejecutar en el programa.

```
1+1 # Suma dos dígitos  
## [1] 2  
5*7 # 5*7  
## [1] 35  
c('a','b','c') # Caracter  
## [1] "a" "b" "c"
```

```

1:7          # Entero
## [1] 1 2 3 4 5 6 7

40<80        # Valor Lógico
## [1] TRUE

2+2 == 5      # Valor Lógico
## [1] FALSE

T == TRUE     # T expresión corta de verdadero
## [1] TRUE

x <- 24       # Asignación de valor 24 a la variable x para su uso posterior (OBJETO)
x/2           # Uso posterior de variable u objeto x
## [1] 12

x             # Imprime en pantalla el valor de la variable u objeto
## [1] 24

x <- TRUE      # Asigna el valor Lógico TRUE a la variable x OJO: x toma el último
valor que se le asigna
x
## [1] TRUE

sum (10,20,30)  # Función suma
## [1] 60

rep('R', times=3) # Repite la letra R el número de veces que se indica
## [1] "R" "R" "R"

sqrt(9)        # Raíz cuadrada de 9
## [1] 3

```

Todas estas operaciones han quedado en la consola, pero no están almacenadas. A continuación hablaremos de la principal característica de *R*, que es su uso de objetos.

## d. Objetos

*R* es un lenguaje de programación por objetos. Por lo cual, vamos a tener objetos a los que se les asigna su contenido. Si usamos una flecha “<-” o “->” le estamos asignando algo al objeto que apunta la flecha. También podemos utilizar el símbolo “=”, pero no se aconseja porque puede dar lugar a malos entendidos.

A continuación se detallan algunos tipos de objetos y cómo podemos acceder a sus elementos utilizando “[ ]” y del mismo modo,

utilizando corchetes, podemos asignar valores dentro de los objetos. El objeto hoy puede ser utilizado para hacer operaciones y para volver a él.

```
y <- c(2,4,6)      # Vector numérico
y <- c("Primaria", "Secundaria") # Vector caracteres
y[2]                  # Acceder al segundo valor del vector y
## [1] "Secundaria"
y[3] <- 'Preparatoria y más' # Asigna valor a la tercera componente del vector
```

Otro elemento es que los objetos tienen características y propiedades. En los siguientes códigos vemos cómo podemos nombrar los elementos de un objeto y cómo podemos saber la clase del objeto que estamos trabajando:

```
sex <- 1:2          # Asigna a la variable sex los valores 1 y 2
names(sex) <- c("Femenino", "Masculino") # Asigna nombres al vector de elementos sexo
sex[2]              # Segundo elemento del vector sex

## Masculino
##      2

z <- c(0, y, 5)    # Concatena escalares y vectores
z

## [1] "0"           "Primaria"        "Secundaria"
## [4] "Preparatoria y más" "5"

w <- vector('numeric', length=10) # Función vector
class(w)

## [1] "numeric"
```

Además de estos objetos tan sencillos, podemos utilizar objetos de tipo lista, matrices y *dataframes*. Las bases de datos de la ENOE, con las que se trabaja en este libro, se importan en este formato. Sobre esta importación ahondaremos en la siguiente sección.

## C. Importando la Encuesta a un ambiente de R

La base de datos de la ENOE era publicada en .dbf hasta hace un par de años. Hoy en día se presenta en cuatro formatos:

- .csv, formato separado por comas, que además se considera como dato abierto.
- .dta, formato compatible con STATA

- .sav, formato compatible con SPSS
- .dbf, formato compatible con dBase

Vamos a cargar los paquetes o instalarlos si es necesario. Puedes copiar los comandos en tu consola o en el *script* de *RStudio*. Después de “##” se presenta lo que se obtendría en la consola como resultado, por lo que se pueden ir replicando cada uno de los pasos.

```
if (!require("pacman")) install.packages("pacman") # instala pacman si se requiere
## Loading required package: pacman
pacman::p_load(tidyverse, haven, sjlabelled) #carga Los paquetes necesarios para este
capítulo
```

Para importar la base desde su formato publicado en Stata, es decir, los archivos con extensión “.dta”, el proceso es muy parecido al de las bases de datos en formato compatible con SPSS. La ventaja de importarla desde este formato es que la labor de etiquetado elaborado por el INEGI no se pierde.

Importamos a nuestro ambiente las cinco bases correspondientes a la encuesta, tal como las publica el INEGI, o se pueden descargar de <<https://tinyurl.com/Enoet319-dta>>. Una vez descargadas, se importan con los siguientes comandos.<sup>10</sup>

```
sdemt319 <- read_dta("./2019_trim3_dta/SDEMT319.dta")
coe1t319 <- read_dta("./2019_trim3_dta/COE1T319.dta")
coe2t319 <- read_dta("./2019_trim3_dta/COE2T319.dta")
hogt319 <- read_dta("./2019_trim3_dta/HOGT319.dta")
vivt319 <- read_dta("./2019_trim3_dta/VIVT319.dta")
```

Cuando vemos lo siguiente “./2019\_trim3\_dta/SDEMT319.dta”, nos referimos a la ruta dónde está guardada nuestra base de datos. Podemos cambiar esta ruta estableciendo claramente la ubicación de los archivos. Sustituyendo donde está “./” por las carpetas de tu máquina. O bien, puedes establecer el directorio donde esté la carpeta en “Session → Set Working Directory → Choose Directory ...” y ahí eliges la carpeta. Esto también se discute en el anexo 1, donde explicamos también el proceso de establecer directorios.

---

<sup>10</sup> Todas las bases fueron descargadas en febrero de 2020 del sitio oficial del INEGI.

Cada base es un objeto en nuestro ambiente. A diferencia de otros programas, podemos tener todas las bases en una misma sesión con el uso de *R*. La importación es bastante sencilla: entre paréntesis está la ruta del archivo y puedes sustituir de acuerdo con la máquina de trabajo. También se puede hacer desde el menú de *RStudio*: “File → Import dataset → from Stata...” y seguir las instrucciones del menú.

## D. Fusión de las bases de datos

Muchas bases de datos están organizadas en varias tablas. La ventaja de la programación por objetos de *R* es que nos permite tener las bases cargadas en nuestro ambiente y llamarlas y juntarlas cuando sea necesario.

Es importante tener claro que las bases de datos pueden tener diferentes dimensiones porque la ENOE es una base de datos de unidad heterogénea, como revisamos en la sección B de este capítulo. Podemos verificar las dimensiones de cada base de datos con el comando “dim()”, que nos dice el número de filas (observaciones) y columnas (variables) para cada base de datos:

```
dim(vivt319) # cada renglón es una vivienda
## [1] 126821      20
dim(hogt319) # cada renglón es un hogar
## [1] 127573      31
dim(sdemt319) # cada renglón es una persona
## [1] 405449      104
dim(coet319) # cada renglón es una persona 12+ y más
## [1] 322363      169
dim(coe2t319) # cada renglón es una persona 12+ y más
## [1] 322363      70
```

Para juntar bases usamos el comando “merge()”. Como argumentos ponemos los nombres de los objetos donde hemos almacenado las bases. Sólo se pueden fusionar dos bases por comando. En “by” se pone el identificador único correspondiente a la variable o variables que forman la llave única, entrecomillado. Cuando se mezclan bases del mismo nivel de análisis, el identificador será igual en ambas bases. Cuando se incorpora información de bases de distinto nivel, debemos escoger la de la base de mayor nivel. Por ejemplo, sabemos que a una vivienda corresponde más de un hogar, por lo tanto, se utiliza el identificador de vivienda para pegar la información con la de hogar.

De acuerdo con el INEGI, los identificadores son como siguen:

- Vivienda {vivienda} es “cd\_a”, “ent”, “con”, “v\_sel”
- Hogares {hogares} es “cd\_a”, “ent”, “con”, “v\_sel”, “n\_hog”, “h\_mud”
- Demográfico {individuos} es “cd\_a”, “ent”, “con”, “v\_sel”, “n\_hog”, “h\_mud”, “n\_ren”
- Cuestionario 1 {individuos} es “cd\_a”, “ent”, “con”, “v\_sel”, “n\_hog”, “h\_mud”, “n\_ren”
- Cuestionario 2 {individuos} es “cd\_a”, “ent”, “con”, “v\_sel”, “n\_hog”, “h\_mud”, “n\_ren”.

Podemos declarar estas llaves en un objeto tipo de vector de caracteres:

```
idvivienda<-c("cd_a", "ent", "con", "v_sel")
idhogar<-c("cd_a", "ent", "con", "v_sel", "n_hog", "h_mud")
idpersona<-c("cd_a", "ent", "con", "v_sel", "n_hog", "h_mud", "n_ren")
```

Cada una de las llaves es un objeto vector de caracteres. No hemos hecho ningún cambio —aún— en las bases de datos.

### a. Fusionado uno a uno<sup>11</sup>

Iniciaremos pegando el cuestionario de ocupación de empleo, que está dividido en dos tablas o bases de datos. Estos tienen el mismo nivel de análisis y la misma cantidad de observaciones, por lo que es más sencillo.

```
coet319<-merge(coe1t319,coe2t319, by=idpersona)
dim(coet319) # hoy tenemos más variables
## [1] 322363    232
```

Un truco aprendido con el tiempo es que hay variables que tienen nombres iguales en las bases, por lo que renombraremos una variable para no confundirnos con las variables del sociodemográfico. En los

---

<sup>11</sup> El fusionado también se puede hacer, en consonancia de *tidyverse* con “join” <<https://dplyr.tidyverse.org/reference/join.html>>. Aquí usamos el comando de base “merge” porque nos parece más sencillo por ser una única función.

siguientes códigos, renombramos las variables p1 y p3 de nuestro objeto coet319 con los nombres “plcoe” y “p3coe”, respectivamente.

```
coet319<- coet319 %>%
  rename(p1coe=p1,
        p3coe=p3)
```

En los códigos más recientes de *R*, se puede observar el operador *pipe* (“tubería”) %>% (Ctrl+Shift+M).

El operador *pipe* %>% del paquete *dplyr* importa este operador de otro paquete (magrittr). Este operador le permite canalizar la salida de una función a la entrada de otra función. En lugar de funciones de anidamiento (lectura desde adentro hacia afuera), la idea de la tubería es leer las funciones de izquierda a derecha, tal como lo mostramos en el código anterior. La operación de renombramiento se hace dentro del objeto coet319. A lo largo del libro iremos mostrando varias maneras de utilizarlo.

Podemos revisar los nombres de la base de datos fusionada con el comando “names()”.

```
names(coet319)
## [ 1] "cd_a"          "ent"           "con"           "v_sel"         "n_hog"
## [ 6] "h_mud"         "n_ren"         "r_def"         "upm.x"        "d_sem.x"
## [11] "n_pro_viv.x"   "n_ent.x"      "per.x"        "eda.x"        "n_inf.x"
## [16] "picoe"         "p1a1"          "p1a2"          "p1a3"         "p1b"
## [21] "p1c"            "p1d"           "p1e"           "p2_1"          "p2_2"
## [26] "p2_3"           "p2_4"          "p2_9"          "p2a_dia"      "p2a_sem"
## [31] "p2a_mes"        "p2a_anio"     "p2b_dia"      "p2b_sem"      "p2b_mes"
## [36] "p2b_anio"       "p2b"           "p2c"           "p2d1"         "p2d2"
## [41] "p2d3"           "p2d4"          "p2d5"          "p2d6"         "p2d7"
## [46] "p2d8"           "p2d9"          "p2d10"         "p2d11"        "p2d99"
## [51] "p2e"             "p2f"            "p2g1"          "p2g2"         "p2h1"
## [56] "p2h2"           "p2h3"          "p2h4"          "p2h9"         "p2i"
## [61] "p2j"             "p2k_anio"     "p2k_mes"      "p2k"          "p3coe"
## [66] "p3a"             "p3b"           "p3c1"          "p3c2"         "p3c3"
## [71] "p3c4"           "p3c9"          "p3d"           "p3e"          "p3f1"
## [76] "p3f2"           "p3g1_1"        "p3g1_2"        "p3g2_1"        "p3g2_2"
## [81] "p3g3_1"          "p3g3_2"        "p3g4_1"        "p3g4_2"        "p3g9"
## [86] "p3g_tot"         "p3h"           "p3i"           "p3j1"         "p3j2"
## [91] "p3k1"           "p3k2"          "p3k3"          "p3k4"         "p3k5"
## [96] "p3k9"           "p3l"           "p4"            "p4_1"         "p4_2"
## [101] "p4_3"           "p4a"           "p4a_1"         "p4b"          "p4c"
## [106] "p4d1"           "p4d2"          "p4d3"          "p4e"          "p4f"
## [111] "p4g"             "p4h"           "p4i"           "p4i_1"        "p5"
## [116] "p5a"             "p5b_hlu"       "p5b_mlu"       "p5b_hma"      "p5b_mma"
## [121] "p5b_hmi"        "p5b_mmi"       "p5b_jhu"       "p5b_mju"      "p5b_hvi"
## [126] "p5b_mv1"        "p5b_hsa"       "p5b_msa"       "p5b_hdo"      "p5b_mdo"
## [131] "p5b_thrs"        "p5b_tdia"      "p5c"           "p5d1"         "p5d_hlu"
## [136] "p5d_mlu"         "p5d_hma"       "p5d_mma"       "p5d_hmi"      "p5d_mmi"
## [141] "p5d_hju"         "p5d_mju"       "p5d_hvi"       "p5d_mv1"      "p5d_hsa"
## [146] "p5d_msa"         "p5d_hdo"       "p5d_mdo"       "p5d_thrs"     "p5d_tdia"
```

```

## [151] "p5e"          "p5f1"         "p5f2"         "p5f3"         "p5f4"
## [156] "p5f5"          "p5f6"          "p5f7"          "p5f8"          "p5f9"
## [161] "p5f10"         "p5f11"         "p5f12"         "p5f13"         "p5f14"
## [166] "p5f15"         "p5f99"         "ur.x"          "fac.x"         "upm.y"
## [171] "d_sem.y"       "n_pro_viv.y"  "n_ent.y"      "per.y"         "eda.y"
## [176] "n_inf.y"       "p6_1"          "p6_2"          "p6_3"          "p6_4"
## [181] "p6_5"          "p6_6"          "p6_7"          "p6_8"          "p6_9"
## [186] "p6_10"         "p6_99"         "p6a1"          "p6a2"          "p6a3"
## [191] "p6a4"          "p6a9"          "p6b1"          "p6b2"          "p6c"
## [196] "p6d"           "p7"            "p7a"           "p7b"           "p7c"
## [201] "p8_1"          "p8_2"          "p8_3"          "p8_4"          "p8_9"
## [206] "p8a"           "p9_1"          "p9_h1"         "p9_m1"         "p9_2"
## [211] "p9_h2"         "p9_m2"         "p9_3"          "p9_h3"         "p9_m3"
## [216] "p9_4"          "p9_h4"         "p9_m4"         "p9_5"          "p9_h5"
## [221] "p9_m5"         "p9_6"          "p9_h6"         "p9_m6"         "p9_7"
## [226] "p9_h7"         "p9_m7"         "p9_8"          "p9_h8"         "p9_m8"
## [231] "ur.y"          "fac.y"

```

¿Qué se puede observar?

- El orden de las variables corresponde al orden en que pusimos las bases en las opciones (primera=x y segunda=y).
- También vemos que las variables que se repetían en ambas bases se repiten en la nueva base, seguida de un punto y una x para lo que proviene de la primera base(x) y con una y lo que proviene de la segunda(y). R dejará las variables intactas y coincidentes, en nuestro caso, porque los nombres de las variables son iguales. R hace esto como precaución ante el posible error de que tengamos alguna variable con un nombre igual y no sea la misma.

Revisemos que las variables “repetidas” son iguales. Esto se puede ver con un tabulado de base con la función “table()”.

```



```

Podemos eliminar las variables después de verificar que son iguales y volver a sus nombres originales con el siguiente código:

```

coet319<-coet319 %>%
  select(-ends_with(".y")) %>% #elimina las variables que terminan en .y
  rename_at(.vars = vars(ends_with(".x")),
            .funs = funs(sub("[.]x$", "", .))) #renombra

```

Vamos a continuar fusionando, hay que tener más cuidado porque tenemos diferentes números de casos y unidades de análisis.

Siempre, en los primeros encuentros con el programa, algunos códigos nos parecen más complejos que otros; por ello, en lo posible, hemos colocado un comentario a la par de cada comando para explicar qué hace. No cómo lo hace, sino qué acción está desarrollando y porqué es necesario correrlo.

En esta primera fusión ya tenemos en un solo objeto la información que venía en las dos bases separadas del Cuestionario de Ocupación y Empleo (COE) del trimestre III de 2019.s

### **b. Tipos de fusionado según información de las bases**

Vamos a fusionar la información de personas mayores de 12 años del COE con el reporte del cuestionario sociodemográfico. Para ello seguiremos usando la misma función, pero nos lleva a explicar que hay un argumento más que no hemos usado con “merge()”.

```
sdemcoet319<-merge(sdemt319, coet319, by=idpersona)
dim(sdemcoet319)
## [1] 322363     320
```

¡La base nueva no tiene a toda la población, sólo la que tiene en la base más pequeña! Tenemos sólo 322 363 de 405 449 personas.

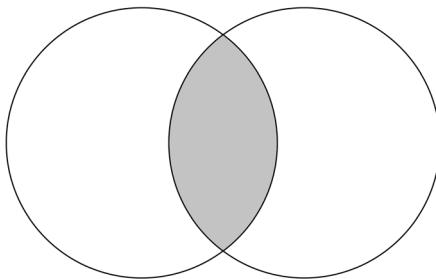
En realidad, hay cuatro formas de fusionar los objetos de tipo *dataframe* de acuerdo con los argumentos “all”, “all.x” y “all.y”.

#### *i) Casos comunes en las dos bases*

Hay opciones preseleccionadas en el programa, por ejemplo, el comando tiene activada la opción “all = FALSE”, que nos deja los casos de ambas bases comunes (tipo una intersección).

```
sdemcoet319<-merge(sdemt319, coet319, by=idpersona, all = FALSE)
dim(sdemcoet319)
## [1] 322363     320
```

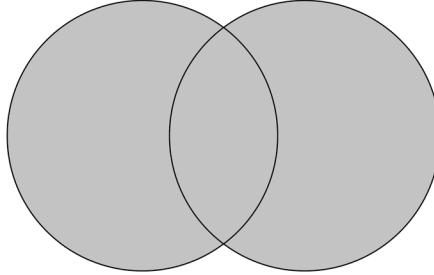
Esto lo podemos resumir con un esquema de teoría de conjuntos. Esta lógica está presente en varios lenguajes de programación y no es exclusiva de *R*.

**Figura II-1. Opción “all=FALSE” con el comando “merge()”**

*ii) Todos los casos en ambas bases*

Si cambiamos la opción “all = TRUE”, se obtienen los comunes a ambas bases y los que aportan cada una de las bases (como una unión).

```
sdemcoet319<-merge(sdemt319,coet319, by=idpersona, all = T)
dim(sdemcoet319)
## [1] 405449    320
```

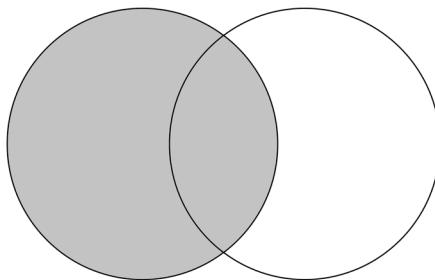
**Figura II-2. Opción “all=TRUE” con el comando “merge()”**

*iii) Casos de la base 1*

Si queremos quedarnos con todos los datos que hay en la primera base, “x”, vamos a usar la opción “all.x = TRUE”.

```
sdemcoet319<-merge(sdemt319,coet319, by=idpersona, all.x = TRUE)
dim(sdemcoet319)
## [1] 405449    320
```

**Figura II-3. Opción “all.x=TRUE” con el comando “merge()”**



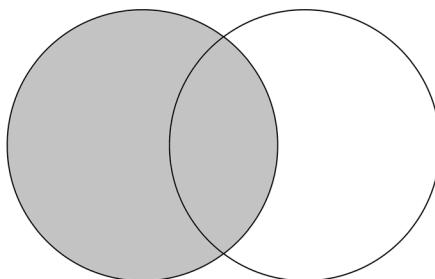
iv) *Casos de la base 2*

Notamos que sí tenemos los datos de toda la población y hay valores perdidos o nulos en las variables aportadas por la base de trabajo.

Si queremos lo contrario, quedarnos con los datos aportados por la segunda base, vamos a usar la opción “all.y=TRUE”. En este caso, como los individuos de la base de ocupación y empleo son un subconjunto del total, es igual a lo que hicimos en el primer caso.

```
sdemcoet319<-merge(sdemt319,coet319, by=idpersona, all.y = TRUE)
dim(sdemcoet319)
## [1] 322363    320
```

**Figura II-4. Opción “all.y=TRUE” con el comando “merge()”**



En realidad, necesitamos la opción donde se tienen todos los casos, por lo que dejaremos lista nuestra base:

```
sdemcoet319<-merge(sdemt319,coet319, by=idpersona, all = TRUE)
dim(sdemcoet319)
## [1] 405449    320
```

Y limpiamos la repetición de variables, lo cual es completamente opcional:

```
sdemcoet319<-sdemcoet319 %>%
  select(-ends_with(".y")) %>% #elimina las variables que terminan en .y
  rename_at(.vars = vars(ends_with(".x")),
            .funs = funs(sub("[.]x$", "", .))) #renombra
```

### c. Fusionando bases de diferente nivel

Hemos fusionado buena parte de nuestros datos. No obstante, en las bases de vivienda y hogar tenemos información que puede ser relevante para nuestro análisis. Por lo que la tarea de fusionado continúa.

Vamos a fusionar la información de vivienda y hogar. A una vivienda pueden corresponder varios hogares, por lo que usaremos el identificador único de vivienda, tal como se explica en la Figura 1-1.

El resultado es una base a nivel de hogar donde las variables de vivienda se repiten para cada hogar que corresponda a la misma. Es decir, al hacer esta fusión nos quedamos con una base donde cada renglón corresponde a la unidad anidada.

```
vivhogt319<-merge(vivt319, hogt319, by=idvivienda)
dim(vivhogt319)
## [1] 127573    47
```

Y limpiamos la repetición de variables, lo cual es completamente opcional:

```
vivhogt319<-vivhogt319 %>%
  select(-ends_with(".y")) %>% #elimina las variables que terminan en .y
  rename_at(.vars = vars(ends_with(".x")),
            .funs = funs(sub("[.]x$", "", .))) #renombra
```

Finalmente, para terminar la fusión, vamos a juntar las bases de individuos y de hogares con vivienda:

```
completat319<-merge(vivhogt319, sdemcoet319, by=idhogar)
```

Y limpiamos la repetición de variables, lo cual es completamente opcional:

```
completat319<-completat319 %>%
  select(-ends_with(".y")) %% #elimina las variables que terminan en .y
  rename_at(.vars = vars(ends_with(".x")),
  .funs = funs(sub("[.]x$", "", .))) #renombra
```

Los documentos de la ENOE establecen que debemos utilizar la información sólo de entrevistas válidas. Es decir, se aconseja eliminar los registros con el campo “r\_def” diferente de “00”, las cuales corresponden a entrevistas incompletas o no logradas, y también eliminar los registros con condición de residencia ausente (“c\_res==2”). Esto se desarrolla con los siguientes códigos:

```
completat319<-completat319 %>%
  filter(r_def==0) %>%
  filter(c_res!=2)
```

Puedes eliminar de tu ambiente el resto de bases y trabajar en este momento sólo con la base completa; ello con el comando “rm”, que remueve objetos. Aquí estamos diciéndole que borre todo objeto que no sea nuestro objeto “completat319”:

```
rm(list=setdiff(ls(), "completat319"))
```

Estamos listos para usar toda la información provista por la ENOE a nivel de individuos. En los capítulos posteriores utilizaremos esta base de datos ya fusionada. La base como objeto en un ambiente de R se puede descargar en <<https://tinyurl.com/completat319-Rdata>>.

## E. Revisión breve de la ENOE

En las siguientes líneas se establecen algunos comandos básicos para revisar una base de datos con la intención de comprender cómo funciona en términos de su contenido. La ENOE es una base muy grande y, como establecimos en la sección anterior, habrá que revisar a profundidad sus documentos metodológicos.

El comando “glimpse()” nos ayuda a “echar un vistazo” el objeto, en este caso, nuestra base de datos. El objeto que queremos revisar se coloca dentro del paréntesis. Como nuestra base es muy

grande y la revisión nos llevaría varias páginas de resultado, agregamos en un corchete una selección de las primeras 10 columnas, que en la lógica de base de datos son variables:

```
glimpse(completat319[,1:10]) # en corchete del Lado derecho podemos ojear columnas 1 a
La 10

## Observations: 397,600
## Variables: 10
## $ cd_a <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ ent <dbl+lbl> 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15...
## $ con <dbl> 40001, 40001, 40001, 40001, 40001, 40001, 40001, 40001, 40001, 40001, 40...
## $ v_sel <dbl+lbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 1, 1, 1, 1, 1, 2, ...
## $ n_hog <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ h_mud <dbl+lbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ loc <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ mun <dbl> 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 31, 31...
## $ est <dbl> 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 10, 10...
## $ est_d <dbl> 201, 201, 201, 201, 201, 201, 201, 201, 201, 201, 201, 201, 201, 201, 201...
```

## F. Selección de casos y de variables

Poco a poco la lógica de *R* se vuelve más comprensible. Es importante considerar que como hay varias maneras de programar, es normal que varios códigos lleven al mismo resultado, lo cual es muy cierto para la selección de casos y variables.

### a. Selección de variables

Para revisar el contenido de un *dataframe* podemos utilizar el formato “basededatos\$var” o bien usar un corchete, que es un operador para entrar en un objeto y para escoger una variable en específico. Estas cuatro formas dan el mismo resultado:

```
<-completat319$t_loc
<-completat319[["t_loc"]] # ¡Ojo con las comillas!
<-completat319[,12]
<-completat319[,"t_loc"]
```

Estas formas son compatibles con la base de *R*, es decir, no necesitamos instalar ninguna paquetería para que se puedan usar.

Para las selecciones, podemos llegar a lo mismo usando el operador *pipe* y el comando “*select()*” de la paquetería *dplyr*.

```
<-completat319 %>%
  select(t_loc)
```

## b. Selección “inversa”

Seleccionar a la inversa podría entenderse como “botar algo”. Se hace seleccionando, pero con precedente de colocar un signo negativo. No funciona con todos los formatos de selección, pero sí con los que presentamos a continuación:

```
x<-completat319 %>%
  select(-t_loc)
x<-completat319[,-12] # "t_Loc" es la variable número 12
rm(x) #rm sólo bota objetos no variables dentro de un objeto
```

Pero con los otros formatos podemos “asignar” valores dentro de un *dataframe* y uno de esos valores puede ser “nada”.

```
completat319$t_loc2<-completat319$t_loc
completat319$t_loc2<-NULL
```

En este sentido, *R* y *RStudio* tienen una curva de aprendizaje mucho más inclinada que otros programas de análisis estadístico. Requieren que el usuario esté mucho más comprometido. En ocasiones debemos comprender en qué forma programó el desarrollador de la librería e, incluso, a veces cómo aprendió quien nos enseñó. Sin duda, en este documento hay mucho de la personalidad de la autora en términos de su programación y quizás haya formas más eficientes de programar y generar resultados. Este documento no pretende tener los mejores códigos, pero sí documentar el trabajo realizado para poder utilizar una base de datos específica en México.

## c. Subconjuntos de datos

Rara vez utilizamos una base de datos completa o queremos hacer operaciones completas con ellas. En general, el objetivo es estudiar a los trabajadores o a las personas en edad de trabajar. Como ya revisamos en el capítulo I, la ENOE nos brinda información representativa para hacer estudios de ciudades específicas.

En general, podemos seleccionar observaciones o filas. Como nuestra base de datos es muy grande, guardaremos el filtro o selección en un objeto nuevo llamado “subset1”.

```
subset1<-completat319[completat319$t_loc==4,]
```

También podemos seleccionar columnas:

```
subset2<- completat319[, c("n_ent", "n_ren", "ing_x_hrs")]
```

Y en este formato de la “base” del programa, podemos combinar los dos tipos de selección:

```
subset3<- completat319[(completat319$t_loc==4 & completat319$sex==1),
c("n_ent", "n_ren", "ing_x_hrs")]
```

Con la paquetería *dplyr*, parte del *tidyverse*, podemos usar los comandos “filter()” y “select()”, utilizando el *pipe* de esta manera:

```
subset4<-completat319 %>%
  filter(t_loc==4 , sex==1) %>%
  select(n_ent, n_ren, ing_x_hrs)
```

#### d. Uso de etiquetas importadas y cómo usarlas

Los objetos que importamos desde su formato .dta son objetos de tipo *dataframe*, pero además tienen una clase que se llama “haven\_labelled”. Con el comando “class()” podemos observar esto.

```
class(completat319$sex)
## [1] "haven_labelled"
```

¿Esto qué significa? Que en la base original hay etiquetas que podemos usar gracias a que las importamos usando el paquete haven. Esto podemos explotarlo muy bien con el comando “as\_label()” que proviene de la paquetería “sjlabelled”. Si quisieramos hacer una tabla de frecuencias de la variable sexo, podríamos hacerla sin y con etiqueta:

```
table(completat319$sex) # sin etiquetas
##
##      1      2
## 192072 205528

table(as_label(completat319$sex)) # con etiquetas
##
## Hombre Mujer
## 192072 205528
```

Podemos hacer un listado de las etiquetas de variables, un tipo “libro de códigos”, que podría ser muy útil usando una función que viene de la librería sjlabelled: “get\_label”. Funciona para toda la base, para columnas o para variables específicas. Nos da la etiqueta de la variable.

```
#print(get_label(completat319)) #todas
print(get_label(completat319[, 5:10])) #de las variables 5 a la 10

##          n_hog      h_mud          loc        mun
## "Número de hogar" "Hogar mudado"    ""       ""
##           est      est_d
##           ""

print(get_label(completat319$clase2)) # de la variable clase2

## [1] "Clasificación de la población en ocupada y desocupada; disponible y no"
```

Con el comando “get\_labels()” obtenemos las etiquetas, pero de los valores de cada una de las variables.

```
#print(get_labels(completat319)) #todas
print(get_labels(completat319[, 5:10])) #de las variables 5 a la 10

## $n_hog
## [1] "Hogar principal" "Hogar adicional" "Hogar adicional" "Hogar adicional"
## [5] "Hogar adicional" "Hogar adicional"
##
## $h_mud
## [1] "Hogar sin cambio"      "Primer cambio de hogar"
## [3] "Segundo cambio de hogar" "Tercer cambio de hogar"
## [5] "Cuarto cambio de hogar"
##
## $loc
## NULL
##
## $mun
## NULL
##
## $est
## NULL
##
## $est_d
## NULL

print(get_labels(completat319$clase2)) # de la variable clase2

## [1] "No aplica"          "Población ocupada"   "Población desocupada"
## [4] "Disponibles"         "No disponibles"
```

Finalmente, cerramos este capítulo con la advertencia de que el usuario debe remitirse a los documentos metodológicos y conceptuales de la encuesta. Algunos elementos conceptuales los revisaremos a lo largo del libro, pero una parte de la actividad propia del

investigador social es revisar por sí mismo —y de manera crítica— los elementos de operacionalización con los que trabaja una base de datos. Es decir, cómo se pasó de un concepto a un indicador. Esta actividad es parte del compromiso ético del uso de datos.

En este capítulo presentamos los paquetes estadísticos a utilizar e hicimos un acercamiento inicial a la ENOE. La hemos fusionado y aprendido a revisarla para obtener información con ella. La importancia de la ENOE es sustantiva en los análisis del mercado de trabajo, por lo que esperamos haber abonado para que más personas se acerquen a esta fuente. En el capítulo siguiente, ahondamos sobre elementos más específicos del mercado de trabajo mexicano y cómo podemos analizarlos utilizando técnicas estadísticas, desde las más descriptivas hasta las que implican modelación de los datos.

### **III. El análisis descriptivo como herramienta de análisis en los mercados de trabajo. El caso de las variables cualitativas**

#### **Introducción**

El análisis descriptivo es una parte esencial en cualquier investigación, incluso en aquellas que quizás no sean de corte tan cuantitativo. Seguramente siempre necesitamos enmarcar nuestro problema de investigación y muchas veces las fuentes, como las encuestas, nos proveen un panorama general, puesto que nos permiten hablar de la población. De ahí que acercarnos a estas generalidades puede ser un paso ineludible y sumamente importante para el inicio investigativo.

El análisis descriptivo en el ámbito de una investigación cuantitativa nunca puede ser omitido. En ocasiones, cuando se piensa que el análisis más complejo es el relevante, se pueden cometer errores por simplemente no haber hecho primero un análisis descriptivo correcto. Además, no siempre lo complejo es mejor. ¿Por qué decir algo de manera compleja cuando se puede hacer simple? Los hallazgos no se miden por la complejidad del método. Como dicen Cortés y Rubalcava (1993, p. 227): “la moda no es buena consejera para la investigación científica; tampoco para escoger entre el arsenal de herramientas estadísticas”. Justo con el advenimiento de los “Grandes Datos”, cada vez más la investigación cuantitativa aparece en la mira, pero con más énfasis en el método que en el resto del proceso investigativo.

Este es el primer capítulo del libro donde se discute el análisis descriptivo. Este capítulo inicia con un repaso del tipo de variables; nos concentraremos en las variables cualitativas y mostramos unos ejemplos usando la Encuesta Nacional de Ocupación y Empleo, par-

tiendo de la fusión que se desarrolló en el capítulo anterior. Posteriormente, revisamos las tablas de doble entrada, cómo se calculan y, sobre todo, cómo se interpretan.

En una tercera parte, introducimos algunos elementos visuales a partir de gráficas de barra y cómo utilizamos escalas de color. Cerramos el capítulo con tres secciones fundamentales. Dos ejemplos de aplicación: el cálculo de tasas de participación y la distribución del empleo de acuerdo con la estructura y el sexo del trabajador. Y, finalmente, realizamos un ejercicio de uso de datos ponderados para replicar algunos de los tabulados publicados por el INEGI.

De esta manera, el objetivo de este capítulo no es sólo introducirnos al análisis descriptivo de variables cualitativas, sino también acercarnos a elementos importantísimos de la configuración del mercado del trabajo mexicano.

## A. Tipos de variables y escalas de medición

El proceso de operacionalización que nos lleva de un concepto a un indicador medible, en la realidad se ve materializado en una matriz de datos; esa tabla donde cada renglón es un individuo y cada columna una variable y que es justo lo que importamos a nuestro paquete estadístico.<sup>12</sup>

Cada una de nuestras variables, o columnas, tiene propiedades específicas y podremos operar elementos estadísticos de acuerdo con la escala de medición con la que se operacionalizó el concepto. Debemos revisar muy bien el diccionario proporcionado por el INEGI para saber qué tipo de variable es. En general, hay cuatro tipos de variables según la escala de medición: nominales, numéricas, de intervalo y de razón.

En este esquema, las operaciones que se hacen en un nivel también se pueden hacer en el siguiente. Es decir que, por ejemplo, la operación de igualdad se realiza en todos los niveles siguientes; esto lo vamos a ejemplificar usando datos de la ENOE en éste y en el capítulo siguiente.

---

<sup>12</sup> La operacionalización es un tema muy complejo que necesita de mucha lectura. Recomendamos lo expresado por Lazarsfeld (1973).

**Cuadro III-1. Ejemplos de variables en la ENOE**

Escala	Operaciones básicas empíricas	Ejemplos de estadísticos	Ejemplos de variables en la ENOE
Nominal	Determinación de igualdad	Frecuencia o número de casos Moda	Sexo (sex)
Ordinal	Determinación de mayor o menor que	Medidas de posición (percentiles, mediana)	Nivel de instrucción (niv_ins)
Intervalo	Determinación de igualdad entre intervalos o diferencias	Media Desviación estándar	Ingresos mensuales (ingocup)
Razón	Determinación de igualdad de razones	Coeficiente de variación	Horas trabajadas

FUENTE: elaboración propia tomando como base a Stevens (1946).

Pasamos entonces a elementos prácticos. Vamos a cargar los paquetes o instalarlos si es necesario. Cargamos la base que obtuvimos en el capítulo anterior, que se puede descargar del siguiente enlace <<https://tinyurl.com/completat319-Rdata>>.

```
if (!require("pacman")) install.packages("pacman") # instala pacman si se requiere
## Loading required package: pacman

#cargan los paquetes necesarios para la práctica de este capítulo
pacman::p_load(tidyverse, haven, janitor, svyr, RColorBrewer, wesanderson, sjlabelled)
load("completat319.RData") ## Recuerda tenerla descargada en tu carpeta de trabajo
```

Para hacer más sencilla nuestra práctica, vamos a utilizar una base de datos que selecciona un par de variables para este capítulo y el siguiente, por lo que crearemos un objeto llamado “base descriptiva” con algunas variables específicas:

```
base_descriptiva<-completat319 %>%
  select(c(sex, eda, niv_ins, t_loc, #vars socdem
          clase1, clase2, clase3, # inserción laboral
          rama, c_ocu11c, # inserción
          ing7c, ingocup, ing_x_hrs, # ingresos
          fac, est_d, upm)) #diseño muestral

rm(completat319)
```

## a. Variables nominales

En la base de datos, la variable nominal “sex” se captura con “1” para hombres y con “2” para mujeres. Posteriormente se establece una operación de igual y se suman los casos que cumplan con esta condición:

```
base_descriptiva %>%
  count(sex==2) # cuentan los casos que cumplen con la condición "sex==2"

## # A tibble: 2 x 2
##   `sex == 2`     n
##   <lgl>     <int>
## 1 FALSE      192072
## 2 TRUE       205528
```

Esto es a lo que nos referimos con contar frecuencias. Podemos contar casos que cumplan con una operación de igualdad.

Una manera más sencilla de observar lo anterior es con una tabla de frecuencias. Una tabla de frecuencias también se considera una distribución. Una característica de la distribución es que proporciona la totalidad de nuestro universo. Por ello, una tabla de frecuencias debe incluir todos los valores que se obtienen en la variable. En el caso de nuestra variable “sexo”, sólo tenemos dos valores posibles.

Introducimos el comando “tabyl()” del paquete *janitor*. Esta función nos da tablas de frecuencias y es un poco más flexible que el comando “table()” de base. Por ejemplo, ofrece ya la distribución de los datos en forma porcentual para nuestro ejemplo de tabulado.

Para que nuestra tabla aparezca ya etiquetada, nos valdremos del operador “mutate()”. Este comando “cambia” los valores de una variable, ya sea en la misma variable o en una nueva. En este caso, dejaremos el mismo nombre, sólo haremos un cambio de los valores de la variable a sus valores en las etiquetas.

```
base_descriptiva %>% mutate(sex=(as_label(sex))) %>% tabyl(sex)

##     sex      n    percent
##   Hombre 192072  0.4830785
##   Mujer 205528  0.5169215
```

Para ver que esto es una distribución de frecuencias, sería muy útil ver la proporción total. Ello se realiza agregando un elemento más en nuestro código con una “tubería”:

```
base_descriptiva %>% mutate(sex=(as_label(sex))) %>% tabyl(sex) %>% adorn_totals()  
## #> #> sex n percent  
## #> Hombre 192072 0.4830785  
## #> Mujer 205528 0.5169215  
## #> Total 397600 1.0000000
```

Como vemos, tenemos 397 600 entrevistas válidas. De las cuales, 51.69% recopilaron información de mujeres y, el resto, de hombres. Al ser una distribución y al sumar los porcentajes de ambas categorías, llegamos a la unidad. La moda es el dato que más se repite; en este caso, aquí la categoría que más se repite es “Mujer”.

Dado que, según el tipo de variables podremos hacer operaciones diferentes, habrá que tener claro el tipo de variable de acuerdo a cómo se recolectó. Pero una cosa es la naturaleza propia de la variable y otra cómo ha sido almacenada.

Es decir, debemos revisar cómo está almacenada la variable en nuestra base de datos. Mucha de esta información se importó desde el formato original y sus etiquetas, por lo que podemos utilizar el comando “as\_label()” para revisar el tipo de variable.

Revisamos algunos tipos de variables:

```
class(base_descriptiva$sex) # variable sin etiqueta  
## [1] "haven_labelled"  
class(as_label(base_descriptiva$sex)) # variable con etiqueta  
## [1] "factor"  
class(as_label(base_descriptiva$niv_ins)) # variable ordinal  
## [1] "factor"  
class(as_label(base_descriptiva$ingocup)) # variable de intervalo/razón  
## [1] "numeric"
```

En general, tendremos variables de “factor” para *R* que podrían ser consideradas como cualitativas y ordinales de acuerdo con su escala de medición. *R* tiene muchas formas de almacenamiento, que iremos revisando poco a poco. Vemos que los ingresos, una variable cuantitativa, tiene el formato “numeric”. Es decir, tenemos formatos de almacenamiento así como tipos de variable de acuerdo con su escala. Y ello a veces depende del conocimiento sobre la base de datos que estamos operando.

Como mostramos con el comando “glimpse()” en el capítulo anterior, podemos revisar una variable en específico:

```
glimpse(base_descriptiva$sex)

##  'haven_labelled' num [1:397600] 1 2 1 1 2 1 1 2 2 2 ...
## - attr(*, "label")= chr "Sexo"
## - attr(*, "labels")= Named num [1:2] 1 2
## ... attr(*, "names")= chr [1:2] "Hombre" "Mujer"
```

Observamos que hay bastante información. Primero aparece un vector numérico que va del 1 al 397 600, es decir, de las dimensiones de nuestra variable. Hay información de la etiqueta de la variable “Sexo”. Luego, otro vector numérico refiere a los niveles de la variable; en este caso, sólo hay dos valores posibles: 1 y 2. A éste se le acompaña con un vector de caracteres que se refiere a las etiquetas (labels) “Hombre” y “Mujer”, por lo que debemos tener cuidado, no todo es lo que parece. En realidad, el 1 y 2 son números, pero están reportando información de una variable de escala nominal, como “Sexo”.

```
base_descriptiva %>% mutate(sex=as_label(sex)) %>% # cambia Los valores de la variable
a sus etiquetas
      tabyl(sex) %>% # para hacer la tabla
      adorn_totals() %>% # añade totales
      adorn_pct_formatting() # nos da porcentaje en Lugar de proporción

## #> #>   sex      n percent
## #>   Hombre 192072    48.3%
## #>   Mujer  205528    51.7%
## #>   Total   397600   100.0%
```

La tubería o *pipe* %>% nos permite ir agregando elementos de manera sencilla a nuestros comandos. En este caso, decimos que dentro del objeto haga el cambio, luego la tabla, que le ponga porcentajes y finalmente que nos dé los totales.

Observa que en realidad hemos solicitado varias operaciones, pero no hemos declarado ningún nuevo objeto. Esto quiere decir que no hicimos cambios en nuestra base de datos.

## b. Variables ordinales

Son variables que dan cuenta de cualidades o condiciones a través de categorías que guardan un orden entre sí. Por ejemplo, el nivel de instrucción. Esta variable es una variable calculada por el INEGI y es una recodificación de las preguntas p13 y p15 del cuestionario sociodemográfico.

Vamos a darle una “ojeada” a esta variable:

```
glimpse(base_descriptiva$niv_ins)
## `haven_labelled` num [1:397600] 4 4 4 4 4 3 1 1 0 4 ...
## - attr(*, "label")= chr "Clasificación de la población ocupada por nivel de instrucción"
## - attr(*, "labels")= Named num [1:6] 0 1 2 3 4 5
## ...- attr(*, "names")= chr [1:6] "No aplica" "Primaria incompleta" "Prrimaria completa"
## "Secundaria completa" ...
```

Hacemos la tabla con las etiquetas de manera similar a la sección anterior:

```
base_descriptiva %>% mutate(niv_ins=as_label(niv_ins)) %>%
  tabyl(niv_ins)

## #> #>   niv_ins      n    percent
## #>   No aplica  28272 0.0711066398
## #>   Primaria incompleta 84722 0.2130835010
## #>   Prrimaria completa 67656 0.1701609658
## #>   Secundaria completa 107439 0.2702188129
## #>   Medio superior y superior 109223 0.2747057344
## #>   No especificado     288 0.0007243461
```

Como se observa en la tabla de frecuencias anterior, podemos hacer operaciones de igualdad. ¿Cuántas personas tienen la primaria incompleta? La tabla nos responde que 84 722, quienes representan 21.3% de la población entrevistada. No obstante, hay una categoría que dice “No aplica”. Como señalamos en el capítulo anterior, es muy importante revisar el material metodológico, así como los cuestionarios.

De acuerdo con el cuestionario sociodemográfico, se establece que las preguntas 12 a 17 sólo se realizan para las personas de 5 años y más; es decir la población en edad escolar.

El “No aplica” siempre proviene de los casos que no llegaron a una pregunta por elementos propios de su población. En este caso, “No aplica” se refiere a las personas menores de 5 años. No tendría sentido alguno medir el nivel de instrucción de personas que no tienen la edad para asistir a la escuela.

Además de “No aplica”, tenemos los valores perdidos o no conocidos. Para concentrarnos en los niveles de instrucción válidos, podemos filtrar estos casos para obtener un tabulado más adecuado de la siguiente forma:

```
base_descriptiva %>%
  filter(eda>5 & niv_ins!=5) %>%
  mutate(niv_ins=as_label(niv_ins)) %>%
  tabyl(niv_ins)

##          niv_ins      n   percent
##        No aplica     0 0.0000000
##    Primaria incompleta 78433 0.2162172
##    Primaria completa 67656 0.1865081
##    Secundaria completa 107439 0.2961784
##  Medio superior y superior 109223 0.3010963
##        No especificado     0 0.0000000
```

El operador “&” permite poner una condición doble y funciona como intersección. Es decir, reportará los casos que cumplen tanto con “eda>5” y “niv\_ins!=5”. En este caso, observamos que el operador “!=” equivale a un “distinto a”.

Para evitar que salgan las categorías sin datos como una fila en nuestra tabla, podemos poner una opción dentro del comando “tabyl()”.

```
base_descriptiva %>%
  filter(eda>5 & niv_ins!=5) %>%
  mutate(niv_ins=as_label(niv_ins)) %>%
  tabyl(niv_ins, show_missing_levels=F ) %>% # elimina los valores 0
  adorn_totals()

##          niv_ins      n   percent
##    Primaria incompleta 78433 0.2162172
##    Primaria completa 67656 0.1865081
##    Secundaria completa 107439 0.2961784
##  Medio superior y superior 109223 0.3010963
##          Total 362751 1.0000000
```

En el Cuadro III-1 señalamos que este tipo de variables permiten, además de las operaciones de igualdad, la determinación de “mayor que” o de “menor que”. Por ejemplo, si queremos saber quiénes de la encuesta llegaron hasta primaria, este es el valor de “2”. Es decir, necesitamos contar a quienes cumplen con la condición “menor a 3” o “menor o igual a 2”.

```
base_descriptiva %>%  filter(eda>5 & niv_ins!=5) %>%
  count(niv_ins<3)

## # A tibble: 2 x 2
##   `niv_ins < 3`     n
##   <lgls>           <int>
## 1 FALSE            216662
## 2 TRUE             146089
```

De ahí que se diga que hay medidas de posición debido a este orden jerárquico. Así, podemos afirmar que 146 089 tienen primaria o menos, es decir, una proporción de 0.4. En otras palabras, el valor de la posición de “primaria completa” divide a la población un 40% inferior o igual a este valor. Mientras que esta posición también implica que 39% supera la primaria. Las medidas de posición separan la distribución de frecuencias en dos grupos que pueden ser leídos desde los valores mínimos y máximos.

Antes de pasar a las variables numéricas, revisaremos cómo se analizan dos variables nominales u ordinales al mismo tiempo.

## B. Tablas de doble entrada

### a. Cálculo de frecuencias

Las tablas de doble entrada deben su nombre a que en las columnas entran los valores de una variable categórica y en las filas los valores de una segunda. Básicamente, es como hacer un conteo de las combinaciones posibles entre los valores de una variable con la otra.

Por ejemplo, para combinar las dos variables que ya estudiamos, lo podemos hacer con una tabla de doble entrada. Vamos a utilizar los filtros de edad escolar y no especificados. Y dentro del comando “tabyl()” agregaremos como argumento nuestra variable “sex”.

```
base_descriptiva %>%
  filter(edad>5 & niv_insl!=5) %>%
  mutate_at(vars(niv_ins, sex), as_label) %>% #mutate_at cambia varias variables
  tabyl(niv_ins, sex, show_missing_levels=F ) %>% # incluimos sexo
  adorn_totals()
```

	niv_ins	Hombre	Mujer
##	Primaria incompleta	38000	40433
##	Primaria completa	32351	35305
##	Secundaria completa	49608	57831
##	Medio superior y superior	54428	54795
##	Total	174387	188364

Las filas corresponden a las categorías de la variable que introdujimos primero en el comando “tabyl()” y las columnas a las categorías de nuestra segunda variable. También, los resultados muestran que en cada celda confluyen los casos que comparten las mismas características.

```
base_descriptiva %>%
  filter(edad>5 & niv_ins!=5) %>% # mantenemos el filtro
  count(niv_ins==1 & sex==1) # nos da la primera celda

## # A tibble: 2 x 2
##   `niv_ins == 1 & sex == 1`     n
##   <lg1>                      <int>
## 1 FALSE                      324751
## 2 TRUE                       38000
```

## b. Totales y porcentajes

Una vez calculadas las frecuencias de las múltiples combinaciones de nuestras categorías, agregamos los totales y los porcentajes. De nuevo utilizaremos el comando “adorn\_totals()”, pero especificando si agrega el total como una columna o como una fila. En el caso de agregarlo como una columna, se tiene:

```
base_descriptiva %>%
  filter(edad>5 & niv_ins!=5) %>%
  mutate_at(vars(niv_ins, sex), as_label) %>% #mutate_at cambia varias variables
  tabyl(niv_ins, sex, show_missing_levels=F ) %>% # incluimos aquí sexo
  adorn_totals("col")

##          niv_ins Hombre Mujer Total
## Primaria incompleta 38000 40433 78433
## Primaria completa 32351 35305 67656
## Secundaria completa 49608 57831 107439
## Medio superior y superior 54428 54795 109223
```

Si queremos una fila, en lugar de “col” se coloca “row” (fila en inglés). También podemos agregar tanto una fila como una columna de totales introduciendo en el argumento “c(“col”, ”row”)” un vector de caracteres de las dos opciones requeridas.

```
base_descriptiva %>%
  filter(edad>5 & niv_ins!=5) %>%
  mutate_at(vars(niv_ins, sex), as_label) %>% #mutate_at cambia varias variables
  tabyl(niv_ins, sex, show_missing_levels=F ) %>% # incluimos sexo
  adorn_totals(c("col", "row"))

##          niv_ins Hombre Mujer Total
## Primaria incompleta 38000 40433 78433
## Primaria completa 32351 35305 67656
## Secundaria completa 49608 57831 107439
## Medio superior y superior 54428 54795 109223
## Total 174387 188364 362751
```

Vamos a añadir una función más que nos dará los porcentajes, los cuales se pueden calcular de tres formas: para las filas, para las columnas o para el gran total poblacional.

Para las columnas, tenemos el siguiente código y los siguientes resultados:

```
base_descriptiva %>%
  filter(eda>5 & niv_ins!=5) %>%
  mutate_at(vars(niv_ins, sex), as_label) %>% #mutate_at cambia varias variables
    tabyl(niv_ins, sex, show_missing_levels=F ) %>% # incluimos aquí sexo
    adorn_totals(c("col", "row")) %>%
    adorn_percentages("col") %>% # Divide los valores entre el total de la
columna
    adorn_pct_formatting() # Lo vuelve porcentaje

##           niv_ins Hombre Mujer Total
## Primaria incompleta 21.8% 21.5% 21.6%
## Primaria completa 18.6% 18.7% 18.7%
## Secundaria completa 28.4% 30.7% 29.6%
## Medio superior y superior 31.2% 29.1% 30.1%
## Total 100.0% 100.0% 100.0%
```

Para hacer cuadros de distribuciones (que todas sus partes suman 100), los porcentajes son de gran ayuda para la interpretación, sobre todo cuando se comparan poblaciones de categorías de diferente tamaño. Por lo general, queremos que los cuadros den información de dónde están los totales y su 100%, de esta manera quien lee se puede guiar con un porcentaje respecto a qué está leyendo. En este caso, vemos que el 100% es común en la última fila. Ello quiere decir que, por ejemplo, el 21.8% del 100% hombres mayores de 5 años tienen un nivel de instrucción de primaria incompleta.

En este tipo de resultados, podemos comparar claramente las estructuras de los niveles escolares entre hombres y mujeres. Vemos que los hombres tienen ligeramente más participación en los niveles medio y superior. Ello podría dar cuenta de las desigualdades en el acceso educativo por género.

Veamos la diferencia de cómo podemos leer la misma celda, que no cambia su frecuencia, si calculamos los porcentajes a nivel de fila:

```
base_descriptiva %>%
  filter(eda>5 & niv_ins!=5) %>%
  mutate_at(vars(niv_ins, sex), as_label) %>%
    tabyl(niv_ins, sex, show_missing_levels=F ) %>%
    adorn_totals(c("col", "row")) %>%
    adorn_percentages("row") %>% # Divide los valores entre el total de la
fila
    adorn_pct_formatting() # Lo vuelve porcentaje

##           niv_ins Hombre Mujer Total
## Primaria incompleta 48.4% 51.6% 100.0%
## Primaria completa 47.8% 52.2% 100.0%
## Secundaria completa 46.2% 53.8% 100.0%
## Medio superior y superior 49.8% 50.2% 100.0%
## Total 48.1% 51.9% 100.0%
```

La última columna coincide con valores del 100%. Es decir, por ejemplo, el 48.4% del 100% de quienes tienen primaria incompleta son hombres. El valor de la celda no cambió, sólo cómo se calculan los porcentajes.

Finalmente, podemos calcular los porcentajes con referencia a la población total en análisis. Es decir, la celda en la esquina inferior derecha de nuestra tabla original.

```
base_descriptiva %>%
  filter(edad>5 & niv_ins!=5) %>%
  mutate_at(vars(niv_ins, sex), as_label) %>% #mutate_at cambia varias variables
    tabyl(niv_ins, sex, show_missing_levels=F) %>% # incluimos sexo
    adorn_totals(c("col", "row")) %>%
    adorn_percentages("all") %>% # Divide Los valores entre el total de la
  población
    adorn_pct_formatting() # Lo vuelve porcentaje

##           niv_ins Hombre Mujer Total
## Primaria incompleta 10.5% 11.1% 21.6%
## Primaria completa  8.9%  9.7% 18.7%
## Secundaria completa 13.7% 15.9% 29.6%
## Medio superior y superior 15.0% 15.1% 30.1%
##             Total 48.1% 51.9% 100.0%
```

De nuevo, el porcentaje cambia porque toma como referencia a toda la población. Podemos señalar que el 10.5% de la población en edad escolar es hombre y tiene primaria incompleta.

## C. Gráficas de barra

Existen varios tipos de gráficas aplicables a las variables. La visualización de datos podría tener su propio libro. Para quien inicia, recomiendo revisar el algoritmo desarrollado en la siguiente página <<https://www.data-to-viz.com/>>, que establece los gráficos que pueden utilizarse de acuerdo con las variables a analizar y el objetivo que se quiere lograr.

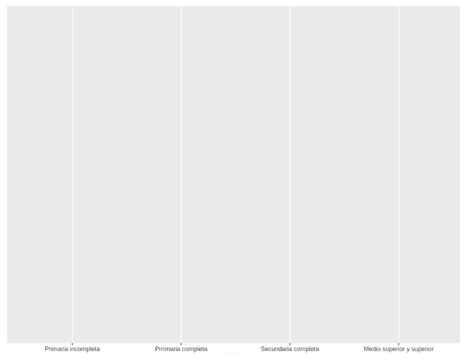
Las gráficas de barra son uno de los medios más importantes para comunicar información. Están diseñadas para variables cualitativas; por lo general mantienen las categorías en uno de los ejes y en el otro se coloca la frecuencia. En esta sección introduciremos, además de un ejemplo de su uso, algunos comandos de *ggplot2*, que es un paquete muy potente para hacer gráficas y quizás el más descargado para su uso en *R*. El prefijo “*gg*” proviene de “Grammar of Graphics” y funciona sintácticamente, de ahí su nombre.

## a. Barras de una variable

Para hacer una gráfica con *ggplot2* se utiliza el comando “*ggplot()*”, cuya función tiene una lógica aditiva. Lo ideal es que iniciemos estableciendo el mapeo estético de nuestra gráfica con el comando “*aes()*”. Lo que creamos es un lienzo donde vamos agregando geometrías que provienen de dibujar nuestros datos.

```
g<-base_descriptiva %>%  
  filter(edad>5 & niv_ins!=5) %>% # filtro de edad válida  
  mutate(niv_ins=as_label(niv_ins)) %>% # cambia valores numéricos a los de las  
etiquetas  
  ggplot(aes(x=niv_ins)) # dibuja nuestro lienzo. En eje de las x irán las categorías  
de la variable niv_ins  
g
```

**Gráfica III-1. Lienzo dibujado por las categorías de la variable “niv\_ins”**

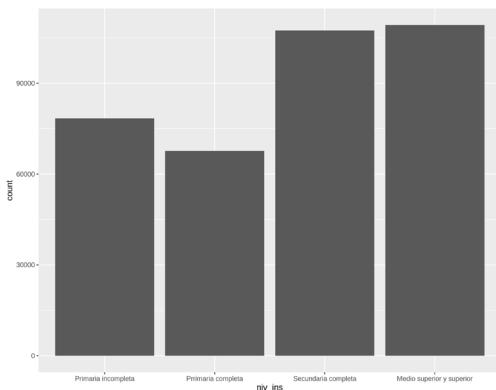


FUENTE: elaboración propia con datos de la ENOE.

Esto nos da un lienzo sobre el cual vamos a pintar y que hemos guardado en un objeto llamado “*g*”, que, al imprimirla en nuestra consola o desde nuestro *script*, produce la gráfica en la ventana de “Plots” de *RStudio*.

Hay un sinfín de geometrías que se adicionan introduciendo un “+” y luego el nombre de la geometría. Para hacer una gráfica de barras, usaremos la geometría “*geom\_bar*”:

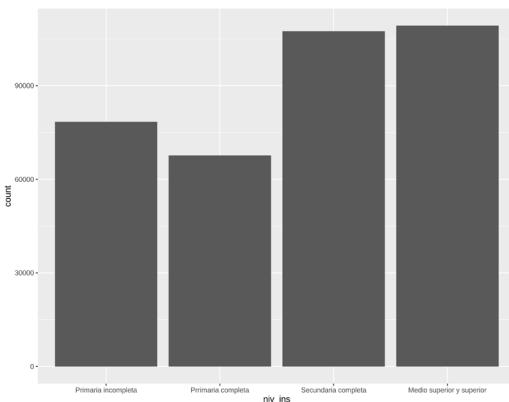
```
g + geom_bar() # dibuja la geometría de barra
```

**Gráfica III-2. Barras de los niveles de escolaridad. Conteos**

FUENTE: elaboración propia con datos de la ENOE.

La gráfica de barras está lista. De forma predeterminada, `geom_bar()` mantiene activado el estadístico “count”, por eso en el eje de las y aparece la palabra *count*.

```
g + geom_bar(stat="count") # dibuja la geometría de barra con conteo
```

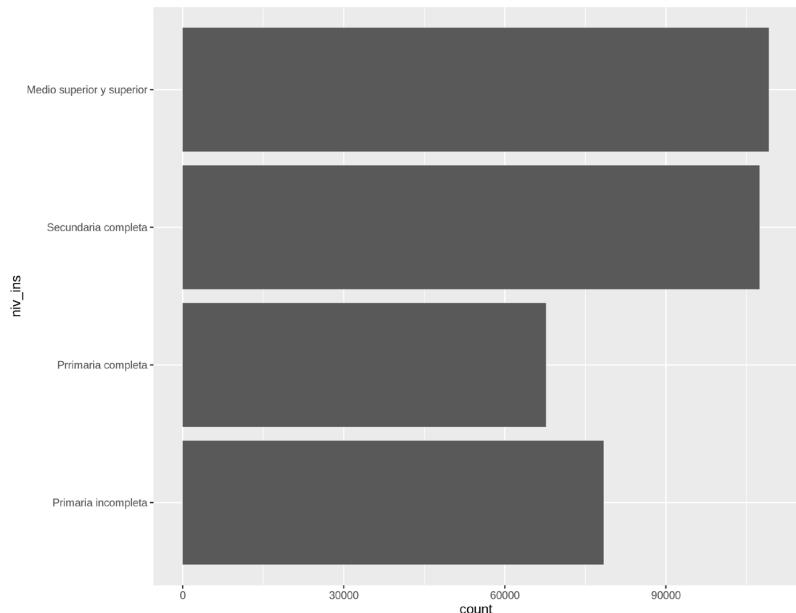
**Gráfica III-3. Barras de los niveles de escolaridad. Conteos.  
Repetición**


FUENTE: elaboración propia con datos de la ENOE.

Para las barras horizontales podemos hacer un giro de nuestras coordenadas sumando un nuevo comando:

```
g + geom_bar(stat="count") + coord_flip()
```

**Gráfica III-4. Barras horizontales de los niveles de escolaridad.**  
**Conteos**

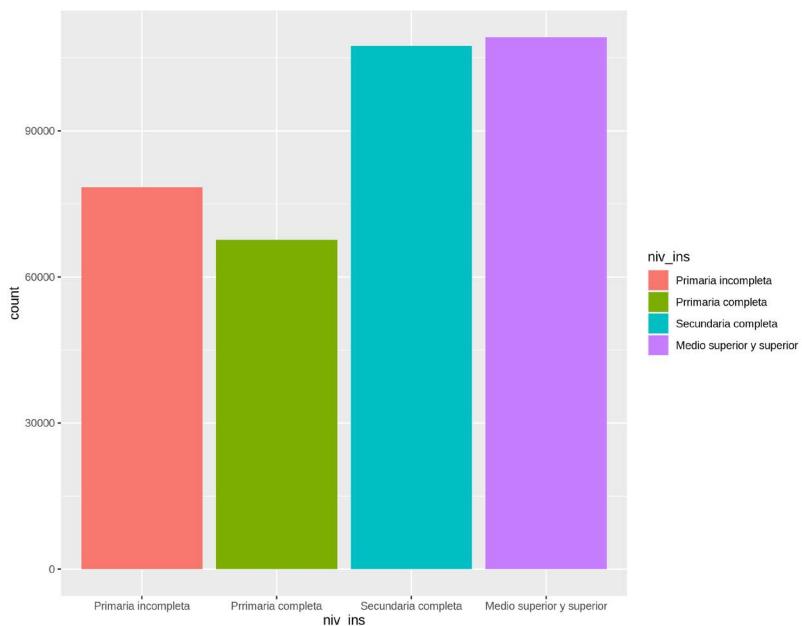


FUENTE: elaboración propia con datos de la ENOE.

Podemos agregar, por ejemplo, un color por cada categoría. Esto lo hacemos como una opción en “aes()”, dentro de “geom\_bar()”. En este tipo de códigos hay que tener mucho cuidado con los paréntesis.

```
g + geom_bar(  
    aes(fill = niv_ins)  
    ) # La barra se rellena usando un color por cada categoría
```

**Gráfica III-5. Barras de los niveles de escolaridad. Conteos. Colores según categoría**

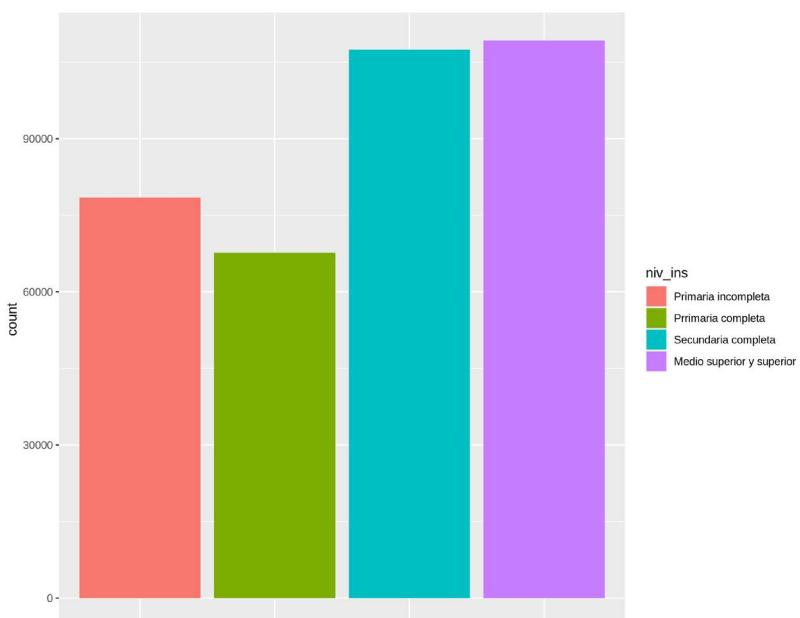


FUENTE: elaboración propia con datos de la ENOE.

Como ya no son necesarias las etiquetas en el eje de las *x*, las eliminamos:

```
g + geom_bar(aes(fill = niv_ins)) + #sumamos una nueva función "theme"
  theme(
    axis.title.x=element_blank(),
    axis.text.x=element_blank(),
    axis.ticks.x=element_blank()
  )
```

**Gráfica III-6. Barras de los niveles de escolaridad. Conteo. Colores según categoría y sin valores en el eje de las x**



FUENTE: elaboración propia con datos de la ENOE.

Poco a poco iremos introduciendo más opciones para las gráficas. Una de ellas es “theme()”, función que nos permite tener control de la apariencia de nuestra gráfica; otra es “element\_blank()”, que por lo general elimina elementos de la gráfica, lo cual puede ser muy útil para hacer gráficas más precisas.

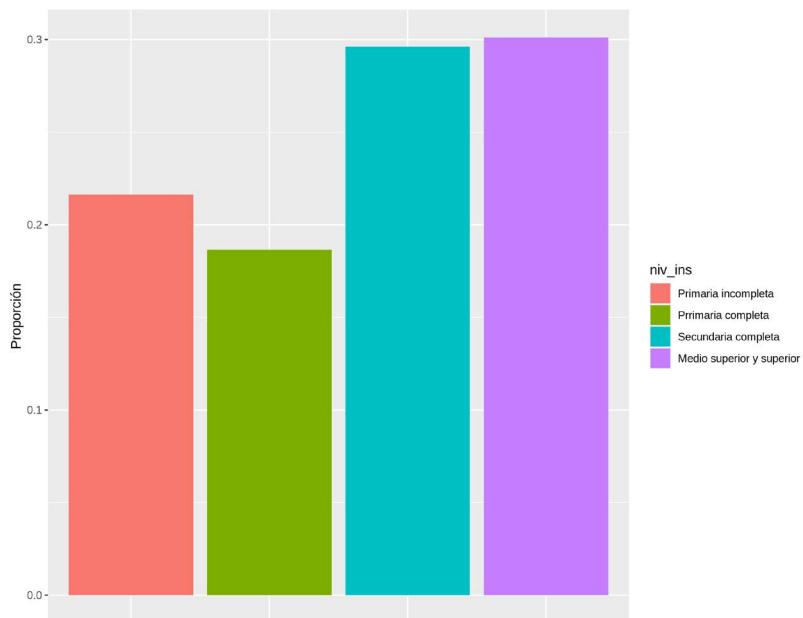
Finalmente, a veces queremos graficar más las proporciones que los conteos en valores absolutos. Esto significa que debemos trabajar en el eje de las y. Si queremos que el conteo que ya tiene se divida entre el total, agregamos para el eje y la definición “y= (.count..)/sum(..count..)” en las opciones geom\_bar asociadas a “aes()” —hemos separado los paréntesis para que se vean muy bien:

```

g + geom_bar(aes(
  fill = niv_ins,
  y=(..count..)/sum(..count..)
) +
ylab("Proporción") + #casi siempre dejamos todas las opciones de "theme" para
el final
  theme(axis.title.x=element_blank(),
  axis.text.x=element_blank(),
  axis.ticks.x=element_blank())

```

**Gráfica III-7. Barras de los niveles de escolaridad. Proporciones.  
Colores según categorías**

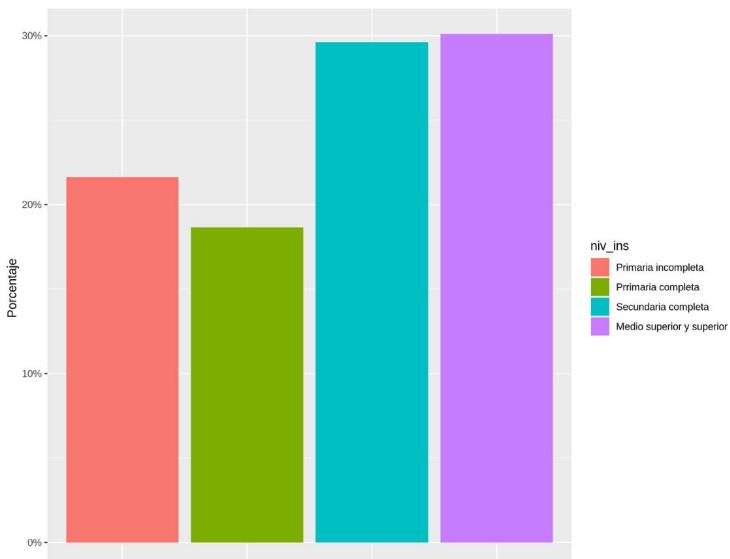


FUENTE: elaboración propia con datos de la ENOE.

Un paso más y podemos agregar los porcentajes. Con esto podemos guardar nuestra gráfica en un objeto y así exportarla para guardarla.

```
g_niv_ins<-g + geom_bar(aes
  (
    fill = niv_ins,
    y=(..count..)/sum(..count..)
  )
) +
ylab("Porcentaje") +
scale_y_continuous(labels=scales::percent) + #casi siempre dejamos todas las
opciones de "theme" para el final
theme(axis.title.x=element_blank(),
axis.text.x=element_blank(),
axis.ticks.x=element_blank()
)
g_niv_ins
```

**Gráfica III-8. Barras de los niveles de escolaridad. Porcentajes. Colores según categoría**



FUENTE: elaboración propia con datos de la ENOE.

Para guardar la gráfica podemos utilizar el comando “`ggsave()`”.

```
ggsave(plot=g_niv_ins, #objeto donde está el gráfico
device="png", # formato del gráfico
filename = "grafico_niv_ins.png") # nombre el archivo de salida

## Saving 5 x 4 in image

# Lo guardará en nuestro directorio y en formato ".png" de imagen
```

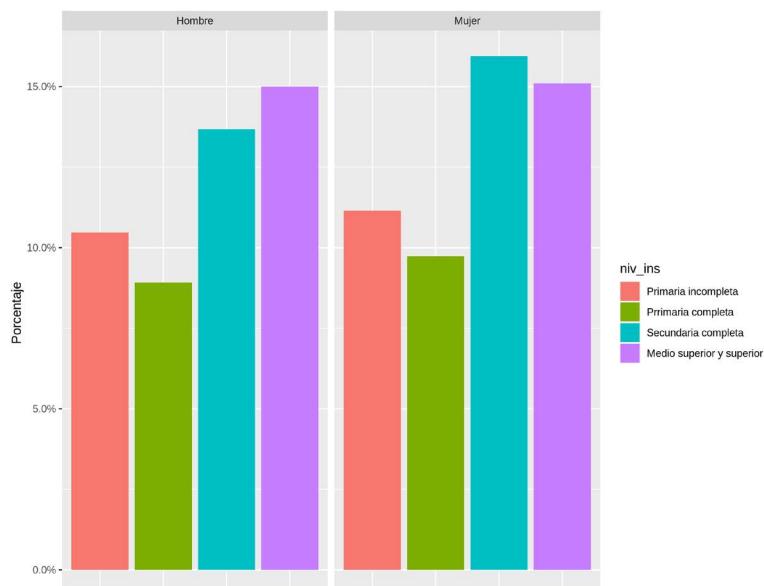
Para más formatos, hay más opciones en la ayuda del comando. En particular el paquete *ggplot()* tiene una curva de aprendizaje bastante inclinada, pero genera gráficas potentes y de todo tipo, así como múltiples extensiones.

### b. Gráficas de barra de dos variables cualitativas

Hay dos maneras de introducir una nueva variable a la gráfica que ya tenemos. Una es replicar cada gráfica que hicimos para cada una de las categorías. Si queremos introducir la variable “sexo”, separamos nuestras poblaciones haciendo un gráfica para cada una. Ello se realiza de manera muy sencilla añadiendo “+ facet\_wrap(~as\_label(sex))”.

```
g_niv_ins + facet_wrap(~ as_label(sex))
```

**Gráfica III-9. Barras de los niveles de escolaridad. Porcentajes. Los paneles identifican las categorías de sexo**



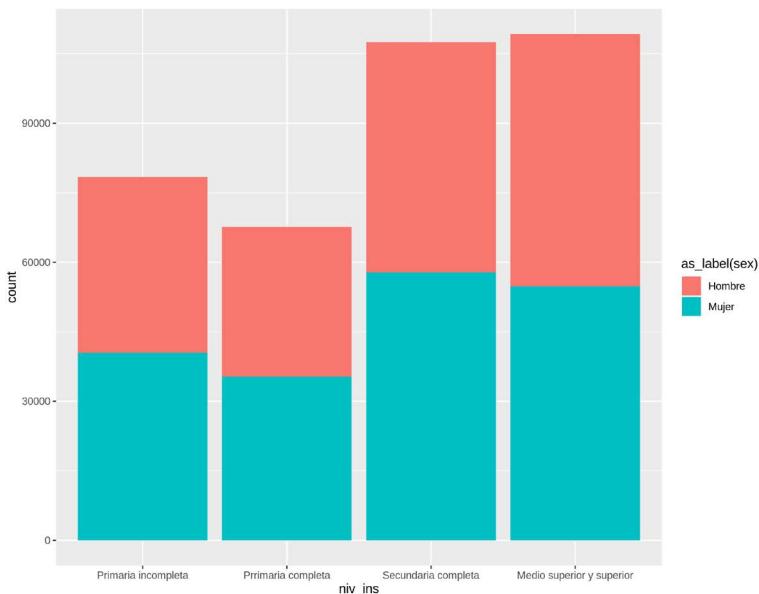
FUENTE: elaboración propia con datos de la ENOE.

Una de las ventajas de *ggplot2* es que la inversión de tiempo para crear una gráfica ahorra mucho tiempo para hacer otras.

Otra forma de incluir una variable es hacer una sola gráfica en la que se distinga por colores a hombres y mujeres. Para esto se sustituye únicamente “*fill=niv\_ins*” por “*fill=as\_label(sex)*” en la gráfica de la sección anterior, sin necesidad de borrar los ejes de las *x*:

```
g + geom_bar(aes(fill = as_label(sex)))
```

**Gráfica III-10. Barras apiladas de los conteos de los niveles de escolaridad según sexo**



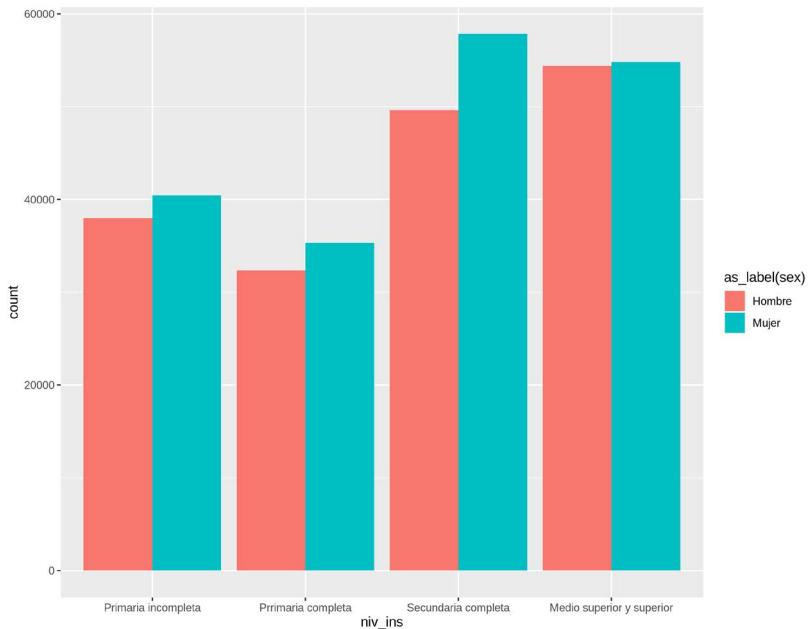
FUENTE: elaboración propia con datos de la ENOE.

Por configuración del comando, las barras aparecen apiladas. Es decir, la barra se mantiene en alto mostrando los conteos para cada categoría de escolaridad; y la barra se “rellena” (de ahí el comando “*fill*”) con los datos que provienen de las categorías de “*sex*”.

Si quisieramos que haya una columna por categoría de sexo, podemos escribir:

```
g + geom_bar(aes(fill = as_label(sex)),
             position="dodge") #pone las categorías Lado a Lado y no apiladas
```

**Gráfica III-11. Barras una contra otra de los conteos de los niveles de escolaridad y sexo**



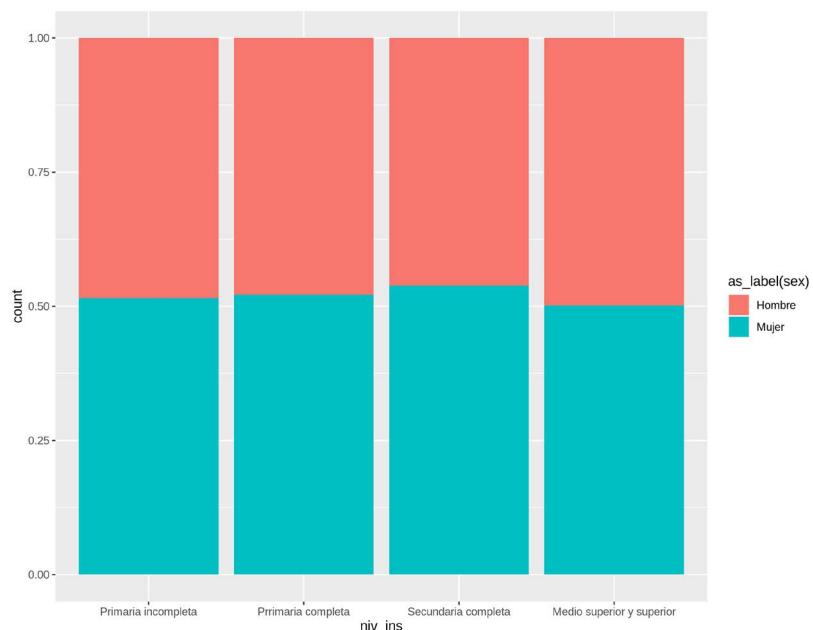
FUENTE: elaboración propia con datos de la ENOE.

Otra manera interesante de observar estas variables es que, estando apiladas, cada categoría de escolaridad se entienda como una unidad:

```
g_bivariado <- g + geom_bar(aes(fill = as_label(sex)),
                           position="fill") # cada categoría "llena" a una unidad
```

g\_bivariado

**Gráfica III-12. Barras apiladas de las proporciones de cada categoría de sexo para cada nivel de escolaridad**



FUENTE: elaboración propia con datos de la ENOE.

### c. Sobre las paletas de colores

Las paletas de color son muy importantes al momento de presentar nuestra información. En *R* hay varios paquetes que permiten crear paletas o bien se pueden crear manualmente.

*RColorBrewer* es un paquete basado en las escalas de color para mapas del sitio <<http://colorbrewer2.org/>>. Este paquete tiene tres tipos de paletas:

1. Las paletas *secuenciales* son adecuadas para los datos ordenados que progresan de menor a mayor. Los pasos de claridad dominan el aspecto de estos esquemas, con colores claros para valores de datos bajos y colores oscuros para valores de datos altos.
2. Las paletas *divergentes* ponen el mismo énfasis en los valores críticos de rango medio y extremos en ambos extremos del rango de datos. La clase

crítica o ruptura en el medio de la leyenda se enfatiza con colores claros y los extremos bajos y altos se enfatizan con colores oscuros que tienen tonos contrastantes.

- Las paletas *cualitativas* no implican diferencias de magnitud entre las clases de leyenda y los tonos se utilizan para crear las principales diferencias visuales entre las clases. Los esquemas cualitativos son los más adecuados para representar datos nominales o categóricos (Traducción libre, Neuwirth, 2014).

La variable “nivel de instrucción”, al ser ordinal, necesita una paleta de color secuencial para que comunique aún mejor los datos. Mientras que “sexo” necesita una cualitativa.

Para ver todas las paletas prediseñadas:

```
display.brewer.all()
```

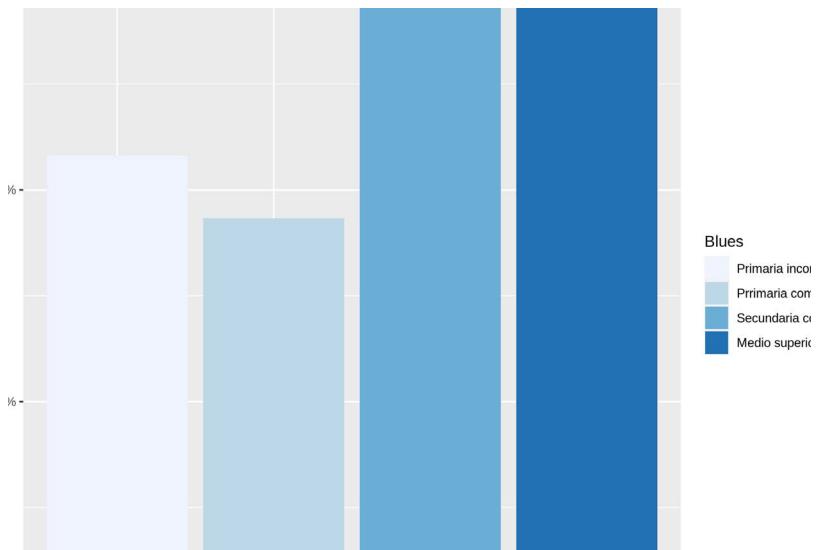
**Figura III-1. Paletas del paquete *RColorBrewer***



*RColorBrewer* tiene ya un comando que se puede utilizar adicional a cualquier objeto que provenga del comando “ggplot” para elegir la paleta secuencial “Blues”:

```
g_niv_ins + scale_fill_brewer("Blues")
```

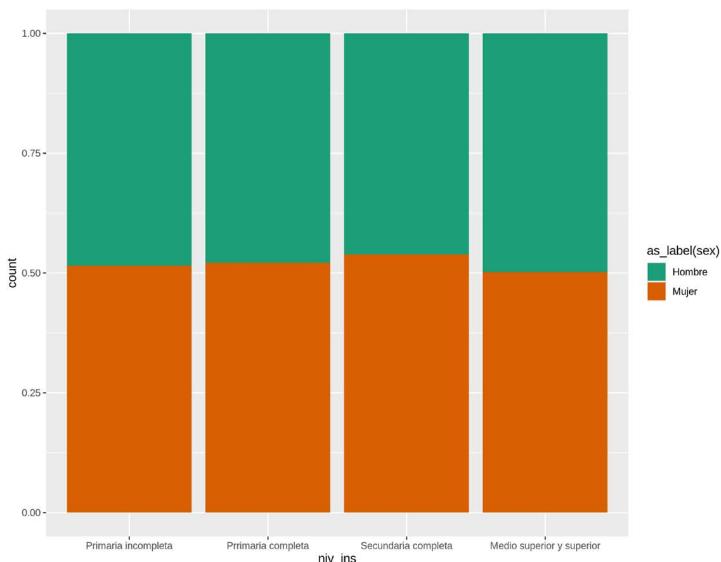
**Gráfica III-13. Gráfica de barras con paleta “Blues”**



FUENTE: elaboración propia con datos de la ENOE.

Para la gráfica de dos variables, donde los colores representan una variable cualitativa, elegimos la paleta cualitativa “Dark2”.

```
col2<- brewer.pal(2, "Dark2") # el 2 es por las dos categorías de sexo  
## Warning in brewer.pal(2, "Dark2"): minimal value for n is 3, returning requested  
palette with 3 different levels  
  
col2 # checa que los colores son hexadecimales. Cualquiera puede hacer sus propias  
paletas  
## [1] "#1B9E77" "#D95F02" "#7570B3"  
g_bivariado + scale_fill_manual(values = col2)
```

**Gráfica III-14. Gráfica de barras apilada con paleta “Dark2”**

FUENTE: elaboración propia con datos de la ENOE.

Se hace la observación de que todas las escalas que coloreamos llevaban el nombre “fill”: “scale\_fill\_brewer” y “scale\_fill\_manual” porque queríamos colorear elementos que colocamos en un argumento de “fill” en el lienzo de nuestra gráfica.

#### **D. Ejemplo a aplicación: estructura productiva y sexo**

Una vez que hemos estudiado lo básico del análisis descriptivo para variables cualitativas y ordinales, corresponde ahora aplicar algunos comandos a variables fundamentales para el estudio de los mercados laborales en México.

Una de las características más documentadas es cómo las mujeres se han concentrado en ramas asociadas a los servicios y los hombres a la industria y agricultura. Esto puede dar lugar a un elemento que se llama “segregación”, que significa que hay distancias entre dos grupos por su adscripción; en este caso, por la condición de ser hombre o mujer (Guzmán, 2009; Kuri, 2014).

Para estudiar esto, utilizaremos las variables “clase2” y “rama\_est2”. Revisaremos sus atributos en los siguientes códigos:

```
attributes(base_descriptiva$clase2)

## $label
## [1] "Clasificación de la población en ocupada y desocupada; disponible y no"
##
## $labels
##      No aplica    Población ocupada Población desocupada
##      0                  1                  2
## Disponibles      No disponibles
##      3                  4
##
## $class
## [1] "haven_labelled"

attributes(base_descriptiva$rama)

## $label
## [1] "Clasificación de la población ocupada por sector de actividad económica"
##
## $labels
##      No aplica          Construcción Industria manufacturera
##      0                  1                  2
## Comercio           Servicios          Otros
##      3                  4                  5
## Agropecuario       No especificado
##      6                  7
##
## $class
## [1] "haven_labelled"
```

Al comparar a hombres y mujeres debemos tomar en cuenta su propia estructura. Las mujeres son menos numéricamente en el trabajo remunerado puesto que muchas se dedican de lleno a labores dentro del hogar. Así que para comparar las propensiones a participar en un sector con respecto a otro, utilizamos un porcentaje dentro de cada grupo por sexo para ver cómo se distribuyen algunas ocupaciones, a pesar de las diferencias de tamaño entre los grupos.

```
ramas<-base_descriptiva %>%
  filter(clase2==1 & eda>14) %>% # seleccionamos a los ocupados
  mutate_at(vars(rama, sex), as_label) %>%
  tabyl(rama, sex, show_missing_levels=F ) %>%
  adorn_totals(c("col", "row")) %>%
  adorn_percentages("col") %>%
  adorn_pct_formatting()

ramas

##                   rama Hombre Mujer Total
## Construcción      12.7%  0.9%  7.9%
## Industria manufacturera 18.0% 16.2% 17.3%
## Comercio          15.5% 24.7% 19.2%
## Servicios         39.4% 54.8% 45.7%
## Otros              1.2%  0.3%  0.8%
## Agropecuario      12.4%  2.6%  8.4%
## No especificado   0.7%  0.5%  0.7%
## Total             100.0% 100.0% 100.0%
```

Es interesante este tipo de análisis porque nos permite ver cómo las mujeres son más proclives a insertarse en sectores terciarios o de servicios, mientras que los hombres tienen mayor participación relativa en la construcción y en la agricultura.

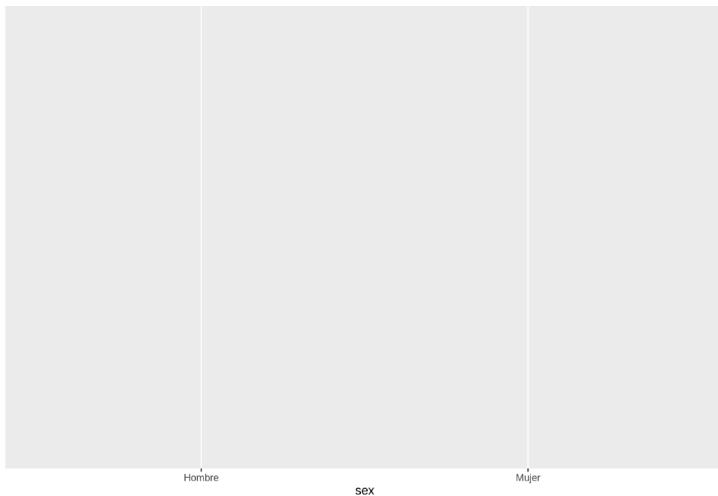
Para hacer una gráfica de barras de estos tabulados también utilizaremos el paquete *ggplot2*. Preparamos primero nuestro lienzo con el mapeo estético y luego colocamos la información que resta.

Como queremos mostrar las diferencias entre la distribución por rama entre hombres y mujeres, lo ideal es que coloquemos nuestra variable de sexo en el “aes()”.

```
g_рамас0<-base_descriptiva %>%
  filter(clase2==1 & eda>14) %>% # seleccionamos a los ocupados
  mutate_at(vars(rama, sex), as_label) %>%
  ggplot(aes(x=sex))
```

g\_рамас0

**Gráfica III-15. Lienzo de la variable “sex”**

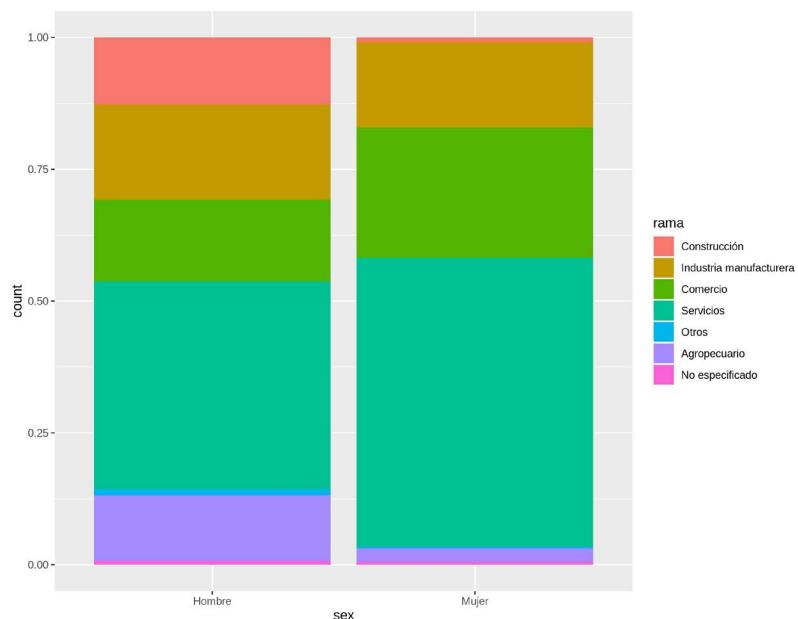


FUENTE: elaboración propia con datos de la ENOE.

Introducimos las ramas:

```
g_рамас0 + geom_bar(aes(fill = rama),
  position="fill") # cada categoría "llena" a una unidad
```

**Gráfica III-16. Barras apiladas de ramas de actividad económica según sexo. Proporción**



FUENTE: elaboración propia con datos de la ENOE.

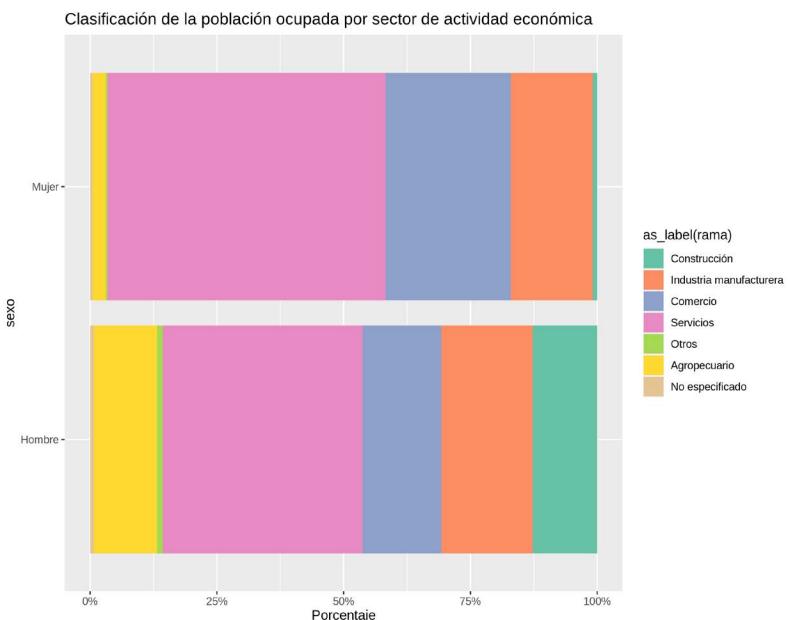
La paleta no parece ser favorecedora, así que vamos a cambiar:

```
col7<-brewer.pal(7, "Set2")
```

Para mejorar la apariencia podemos etiquetar los ejes y cambiar la dirección de las etiquetas del eje de las *x*, así como el cambio en colores:

```
g_ramas0 + geom_bar(aes(fill = as_label(rama)),  
                      position="fill") + # cada categoría "Llena" a una unidad  
                      labs(y="Porcentaje", x="sexo", # escribe las etiquetas del gráfico  
                            title=paste(get_label(base_descriptiva$rama))) + # Truco para titular  
                            scale_y_continuous(labels=scales::percent) + # Pasa porcentaje las escala  
                            scale_fill_manual(values=col7) + # Pasa porcentaje las escala  
                            coord_flip()
```

**Gráfica III-17. Barras apiladas horizontales de ramas de actividad económica según sexo. Porcentajes**



FUENTE: elaboración propia con datos de la ENOE.

Revisemos la parte “`paste(get_label(base_descriptiva$rama))`”. Si corremos esto en nuestra consola queda así:

```
paste(get_label(base_descriptiva$rama))
## [1] "Clasificación de la población ocupada por sector de actividad económica"
```

El comando “`paste()`” pega el valor de un objeto como texto, de ahí que sea muy útil para titular elementos. “`paste()`” pega los valores del objeto que se establecen, en este caso la etiqueta de variable de rama.

Podemos incluso hacer cosas más complicadas:

```
paste(get_label(base_descriptiva$rama), "según", get_label(base_descriptiva$sex), sep=" ")
## [1] "Clasificación de la población ocupada por sector de actividad económica según Sexo"
```

“paste()” pega diferentes valores de objetos, separados por coma. Con el argumento “sep” se separan estos objetos por algún elemento. Si no se quiere separar, se coloca el argumento de la siguiente manera: sep="".

## E. Las tasas de participación económica

Uno de los grandes temas de los mercados de trabajo es la participación en él. ¿Quiénes llegan a estar económicamente activos? ¿Cuál es su perfil? Para ello se utilizan las tasas de participación económica. En general, las tasas o proporciones de participación económica se calculan a nivel quinquenal para la edad y ello muestra el patrón etario de esta participación. Este procedimiento también nos ayudará a recodificar una variable.

Como establecimos al inicio de este capítulo, la variable “eda” tiene una escala de intervalo, o razón, y está almacenada con un formato numérico. Observa a continuación.

```
glimpse(base_descriptiva$eda)  
##  num [1:397600] 70 66 46 43 33 15 11 7 3 60 ...
```

Pero podemos problematizar lo que entendemos por edad. La edad es el tiempo que hemos vivido desde nuestro nacimiento; la escala numérica de año es una escala discreta. La naturaleza de la variable es numérica, que podría ser incluso contada con decimales. Si tuviéramos exactamente la fecha de nacimiento contra la fecha de entrevista, podríamos tener información con precisión menor a un año.

La escala elegida para el registro es de “años cumplidos”. Esta es una escala numérica discreta y podemos recodificar cambiando la escala. Recodificar variables en escalas menos complejas, cuando vamos de abajo hacia arriba, es sencillo (véase Cuadro III-1). Mientras que lo contrario (de arriba hacia abajo) rara vez es posible.

A continuación crearemos una nueva variable que recodifica la edad como quinquenal, es decir, cada cinco años. De nuevo nos auxiliaremos del comando “mutate()” de *dplyr*, pues no sólo cambia valores sino que puede crear nuevas variables:

```
base_descriptiva<-base_descriptiva%>%
  mutate(eda5=cut(eda, # La variable a cortar
                 breaks=seq(0,100, # El rango válido
                            by=5), # El ancho del intervalo
                 include.lowest=T, # incluye el valor más bajo del intervalo
                 right=F)) #el intervalo abierto en la derecha=no
```

Es importante guardar de nuevo el objeto si queremos que los cambios realizados con “`mutate()`” permanezcan en nuestra base de datos. Por ello escribimos “`base_descriptiva <-`” antes de nuestras operaciones. De lo contrario, los pasos realizados no se quedan guardados en el objeto, a pesar de que el programa realice las transformaciones. Al no darle la orden de guardarla, no cambia el objeto en nuestro ambiente.

Veamos cómo quedó esta nueva variable:

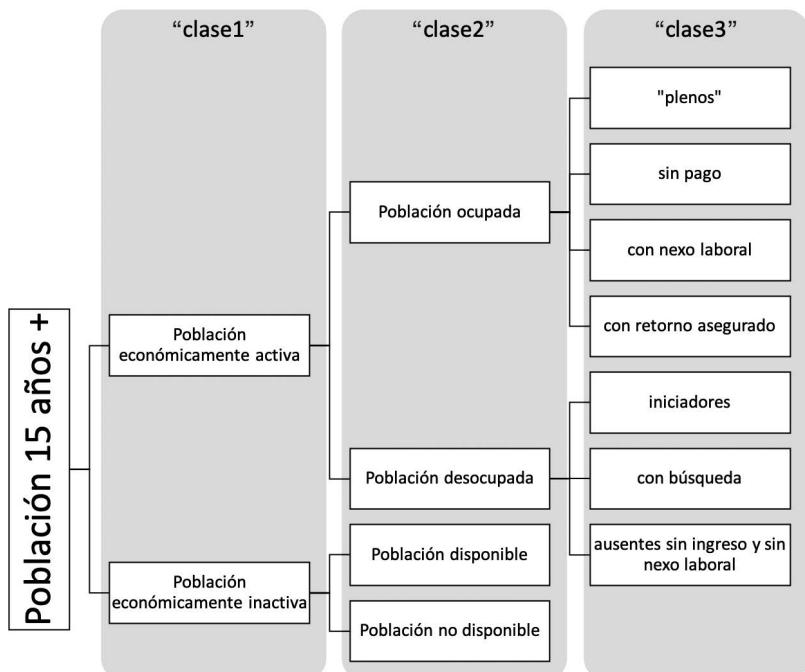
```
base_descriptiva %>% tabyl(eda5)

##      eda5     n   percent
## [0,5)  28272 0.071106640
## [5,10) 33208 0.083521127
## [10,15) 34756 0.087414487
## [15,20) 36491 0.091551811
## [20,25) 33271 0.083679577
## [25,30) 29863 0.075108149
## [30,35) 27682 0.069622736
## [35,40) 26804 0.067414487
## [40,45) 27498 0.068933602
## [45,50) 25874 0.065075453
## [50,55) 23213 0.058382797
## [55,60) 19713 0.049579980
## [60,65) 16177 0.040686620
## [65,70) 12361 0.031089034
## [70,75) 8791 0.022110161
## [75,80) 6200 0.015593561
## [80,85) 3838 0.009652918
## [85,90) 2310 0.005809859
## [90,95) 845 0.002125252
## [95,100] 613 0.001541751
```

Tenemos intervalos semiabiertos de edad. Esto nos permite utilizar esta variable, que era numérica, como una serie de intervalos ordinales. Luego crearemos un nuevo objeto donde irán nuestras participaciones en el mercado de trabajo.

Utilizaremos la variable “`clasel`” del cuestionario sociodemográfico. Esta variable fue construida por el INEGI e identifica a la población no económicamente activa y a la económicamente activa. Es una de la serie de variables que permite identificar diferentes tipos de (no) inserción en los mercados laborales. La lógica de estas variables se puede observar en el siguiente esquema:

**Figura III-2. Clasificación de la población en edad de trabajar según las variables “clase1”, “clase2” y “clase3”**



FUENTE: elaboración propia con la ENOE.

De acuerdo con el INEGI (2019, pp. 12-14), las categorías de “clase1” se refieren a lo siguiente:

Una persona pertenece a la PEA, en términos del mercado laboral, si forma parte del grupo de proveedores u oferentes de servicios laborales, algunos de los cuales han logrado que alguien demande sus servicios, es decir, fueron contratados para desempeñar una actividad económica (ocupados); mientras que otros, aunque aún no lo consiguen (desocupados), están ejerciendo una presión a través de la búsqueda de trabajo, acción que también influye en los mercados laborales.

[...] La población clasificada en esta categoría [PNEA] se refiere a aquella porción de la población no ocupada cuya subsistencia se basa en la transferencia de ingresos monetarios o no monetarios realizada por un familiar o terceras partes, y que además no intenta modificar esa condición de no ocupación involucrándose en el mercado laboral (cosa que los distingue de los

desocupados). Cabe resaltar que se considera que la población clasificada en la categoría de no económicamente activa desempeña un papel relevante tomando en cuenta que realizan actividades que, si bien son ajena al ámbito de la transacción de mercado, no por ello dejan de ser cruciales para el funcionamiento de los hogares y de la sociedad en general: justo porque hay alguien que se encarga de los quehaceres del hogar o de cuidar o atender a los hijos, enfermos o ancianos, alguien más de los integrantes del hogar puede salir a trabajar o a buscar trabajo.

Las etiquetas de la variable “clase1” son bastante largas, sobre todo para tablas y gráficas, y quizás sean preferibles etiquetas más cortas.

```
get_labels(base_descriptiva$clase1)
## [1] "No aplica"
## [2] "Población económicamente activa"
## [3] "Población no económicamente activa"
```

Vamos a crear un par de objetos para almacenar información de etiquetas y después procederemos a establecer las etiquetas. Las nuevas etiquetas deben coincidir en orden con las anteriores.

```
clase1.shortlab<-c("NA", "PEA", "PNEA") # etiquetas a establecer
clase1.longlab<-get_labels(base_descriptiva$clase1) # guardamos las etiquetas anteriores
por si quisieramos volverlas a usar

base_descriptiva$clase1<-set_labels(base_descriptiva$clase1,
                                       labels=clase1.shortlab)

get_labels(base_descriptiva$clase1)
## [1] "NA"    "PEA"   "PNEA"
```

Haremos nuestras tablas. El comando “tabyl()” permite incluir hasta tres variables en análisis. La primera variable quedará en la fila, la segunda en las columnas y la tercera en lista. Por lo general, y para mejorar el formato, preferimos poner la variable con más categorías en las filas. Por tanto, nuestro código queda de la siguiente manera:

```
#clase1 1="PEA", 2= PNEA
tasas<-base_descriptiva %>%
  filter(eda>14 & eda<99) %>%
  mutate_at(vars(sex, clase1), as_label) %>%
  tabyl(edad, clase1, sex, show_missing_levels = F) #tabyl() con tres variables
tasas
```

```
## $Hombre
##      eda5   PEA  PNEA
## [15,20)  7567 10859
## [20,25) 12429  4070
## [25,30) 13391 1198
## [30,35) 12472  592
## [35,40) 12022  517
## [40,45) 12229  604
## [45,50) 11419  697
## [50,55)  9734  959
## [55,60)  7564 1474
## [60,65)  4948 2476
## [65,70)  3077 2619
## [70,75)  1586 2391
## [75,80)   918 1906
## [80,85)   329 1332
## [85,90)   115  865
## [90,95)    23  292
## [95,100]   100 145
##
## $Mujer
##      eda5   PEA  PNEA
## [15,20) 3978 13997
## [20,25) 8180  8592
## [25,30) 9449  5915
## [30,35) 9096  5522
## [35,40) 8917  5348
## [40,45) 9190  5385
## [45,50) 8396  5362
## [50,55) 6956  5564
## [55,60) 4901  5774
## [60,65) 2981  5772
## [65,70) 1581  5084
## [70,75)  811  4003
## [75,80)   391  2993
## [80,85)   139  2038
## [85,90)    48  1282
## [90,95)     7  523
## [95,100]   66  282
```

Algo que debemos tener en cuenta es que hay pocos valores en las últimas celdas, pues muy pocas personas están representando un grupo, por lo cual es posible que las estimaciones no sean adecuadas. Podría ser que esas personas se comportaran de esa manera simplemente por el azar. Hay procesos estadísticos (como el cálculo del coeficiente de variación en el diseño muestral) que permiten una medida de qué tanto podemos hablar de estas categorías.

Antes de calcular nuestras proporciones hasta la población de 85 años, revisemos qué tipo de resultados arroja “tabyl()” con tres variables en el nuevo objeto llamado “tasas”. Este objeto es una lista y podemos imprimir los elementos de una lista utilizando el símbolo “\$” después del nombre del objeto:

```

class(tasas)
## [1] "list"
tasas$Hombre

##      eda5    PEA   PNEA
##  [15,20)  7567 10859
##  [20,25) 12429  4070
##  [25,30) 13391  1108
##  [30,35) 12472  592
##  [35,40) 12022  517
##  [40,45) 12229  604
##  [45,50) 11419  697
##  [50,55)  9734  959
##  [55,60)  7564 1474
##  [60,65)  4948 2476
##  [65,70)  3077 2619
##  [70,75)  1586 2391
##  [75,80)   910 1906
##  [80,85)   329 1332
##  [85,90)   115  865
##  [90,95)    23  292
##  [95,100]   100  145

tasas$Mujer

##      eda5    PEA   PNEA
##  [15,20)  3978 13997
##  [20,25)  8180  8592
##  [25,30)  9449  5915
##  [30,35)  9096  5522
##  [35,40)  8917  5348
##  [40,45)  9190  5385
##  [45,50)  8396  5362
##  [50,55)  6956  5564
##  [55,60)  4901  5774
##  [60,65)  2981  5772
##  [65,70)  1581  5084
##  [70,75)   811  4003
##  [75,80)   391  2993
##  [80,85)   139  2038
##  [85,90)    48  1282
##  [90,95)     7  523
##  [95,100]   66  282

```

Cada elemento de la lista es una pequeña base de datos con los resultados, como se observa con el comando “class”:

```

class(tasas$Hombre)
## [1] "tabbyl"      "data.frame"

```

Una vez que tenemos las frecuencias en una lista calculamos las proporciones. Como queremos calcular la proporción de ocupados dentro de cada grupo quinquenal, siguiendo la misma manera de colocar la información, debemos obtener proporciones de fila. También agregamos los totales.

```

tasas<-base_descriptiva %>%
  filter(edad>14 & edad<=85) %>% #vamos a recortar
  mutate_at(vars(sex, clase1), as_label) %>%
  tabyl(edad, clase1, sex, show_missing_levels = F) %>%
  adorn_percentages("row") # calcula proporciones dentro de cada fila
tasas

## $Hombre
##   edad5      PEA      PNEA
## [15,20) 0.4106697 0.58933029
## [20,25) 0.7533184 0.24668162
## [25,30) 0.9235899 0.07641906
## [30,35) 0.9546846 0.04531537
## [35,40) 0.9587686 0.04123136
## [40,45) 0.9529338 0.04706616
## [45,50) 0.9424728 0.05752724
## [50,55) 0.9183152 0.08968484
## [55,60) 0.8369198 0.16308918
## [60,65) 0.6664871 0.33351293
## [65,70) 0.5492037 0.45979635
## [70,75) 0.3987931 0.60120694
## [75,80) 0.3231534 0.67684659
## [80,85) 0.1980734 0.80192655
## [85,90) 0.1395349 0.86046512
##
## $Mujer
##   edad5      PEA      PNEA
## [15,20) 0.22130737 0.7786926
## [20,25) 0.48771762 0.5122824
## [25,30) 0.61500911 0.3849909
## [30,35) 0.62224655 0.3777535
## [35,40) 0.62509639 0.3749036
## [40,45) 0.63053173 0.3694683
## [45,50) 0.61026312 0.3897369
## [50,55) 0.55559105 0.4444089
## [55,60) 0.45911007 0.5408899
## [60,65) 0.34056895 0.6594311
## [65,70) 0.23720930 0.7627907
## [70,75) 0.16846697 0.8315330
## [75,80) 0.11554374 0.8844563
## [80,85) 0.06384933 0.9361507
## [85,90) 0.05157593 0.9484241

```

Las tasas, por lo general, se grafican con líneas que se forman al juntar puntos del punto medio del intervalo de edad con los valores de las tasas. Para una mejor comparación conviene ponerlas en un solo objeto de tipo *dataframe*, por lo que las juntaremos en una misma “base”.

Para juntarlas, primero identificamos el sexo en una variable y luego reunimos la información con el comando “*rbind()*”. Como se señala en anexos, este comando junta la información sumando renglones. Con esto evitamos tener un objeto tipo lista y obtenemos un solo *dataframe* con la variable sexo.

```

tasas$Hombre<-tasas$Hombre %>%
  mutate(sex="Hombre")
tasas$Mujer<-tasas$Mujer %>%
  mutate(sex="Mujer")
ggtasas<-rbind(tasas$Hombre, tasas$Mujer)

```

Para graficar utilizaremos “ggplot()”, pero con geometría de línea. Primero, establecemos nuestro lienzo. Esta gráfica tendrá en el eje de las  $x$  los intervalos de edad y en el de las  $y$  el valor de las tasas, que son proporciones que no pueden tomar un valor mayor a 1 en cada grupo quinquenal. Definimos en la estética tanto nuestra variable  $x$  como la variable  $y$ , y agrupamos los resultados por sexo. Esto permitirá que la línea quede mejor dibujada:

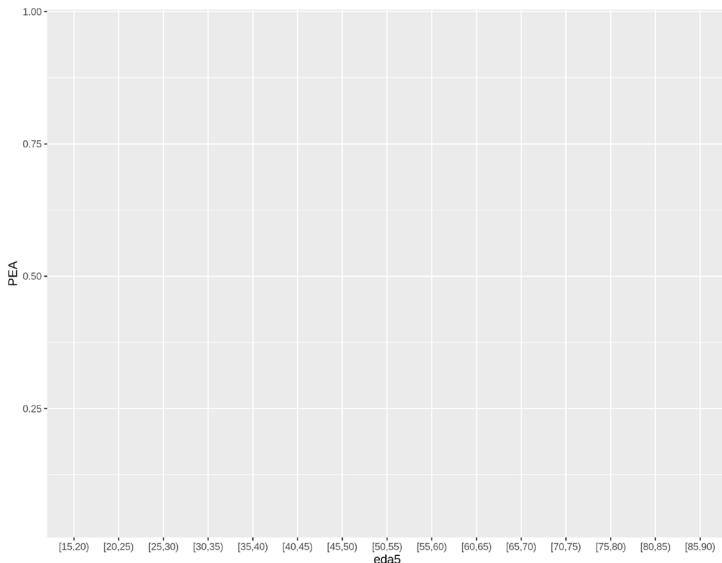
```

g_tasa0<-ggtasas %>%
  ggplot(aes(x=eda5, y=PEA, group = sex))

g_tasa0

```

**Gráfica III-18. Lienzo bivariado. En el eje de las  $x$  tenemos los intervalos de edad y en el de las  $y$  al valor de la población parte de la PEA en la PET**

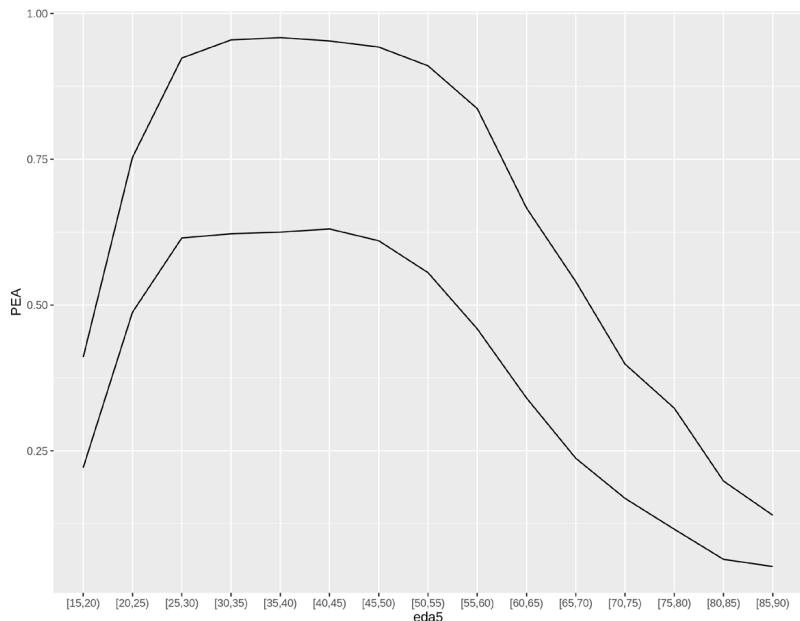


FUENTE: elaboración propia con datos de la ENOE.

Vamos a agregar nuestras líneas que unen los puntos de las tasas:

```
g_tasa0 + geom_line() # dibuja Líneas
```

**Gráfica III-19. Tasas de participación económica en México.  
Trimestre III, 2019**

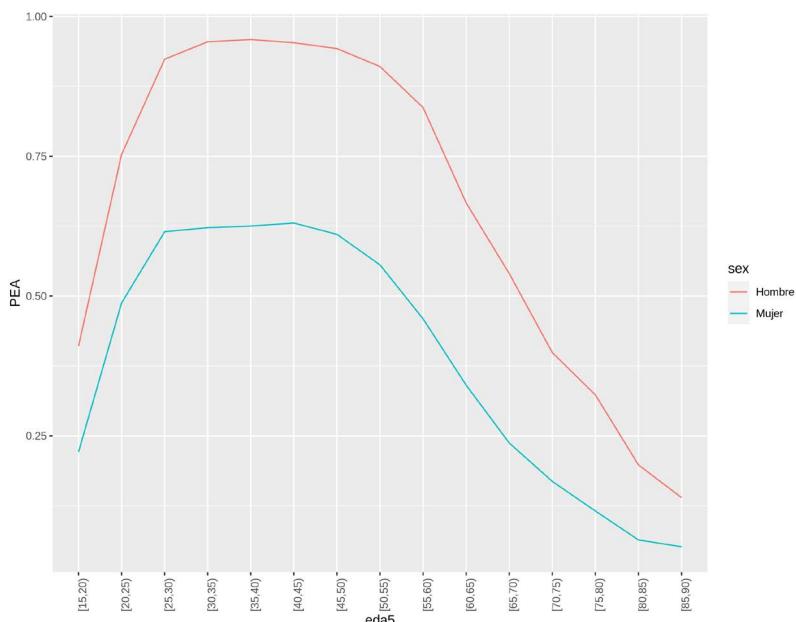


FUENTE: elaboración propia con datos de la ENOE.

Para mejorar la gráfica, pondremos colores para las categorías y le cambiaremos la orientación a las etiquetas del eje de las *x*:

```
g_tasa0 + geom_line(aes(color=sex)) + # dibuja Líneas con colores según categorías de sexo  
theme(axis.text.x = element_text(angle = 90))
```

**Gráfica III-20. Tasas de participación económica en México.  
Trimestre III, 2019. Los colores representan el sexo**



FUENTE: elaboración propia con datos de la ENOE.

Sin duda, la gráfica muestra claramente una menor participación de las mujeres dentro de la PEA. Esto, de nuevo, destaca algunos elementos de la división sexual del trabajo y los roles sexo-genéricos, donde las mujeres realizan más actividades dentro del hogar. Si alguien quiere adentrarse en este tema, hay un cúmulo de estudios sobre ello (i.e: Abramo & Valenzuela, 2005; Cruz & Reyes, 2015; Escoto, 2020; García & Pacheco, 2011).

En términos de procedimiento y procesamiento de la información, esta gráfica, por varios elementos, es diferente a las que habíamos realizado. Primero, no la creamos desde el objeto de la base de datos, sino a partir de los resultados de una operación que almacenamos en otros objetos. Ello implicó algunas operaciones y, sobre todo, un cambio en la unidad de análisis.

En estricto sentido, esta gráfica no es de variables cualitativas. La variable “clase1”, que era cualitativa y de escala nominal para cada individuo entrevistado en edad de trabajar en el objeto “base\_descriptiva”, cambió su naturaleza. Una vez que calculamos la proporción dentro de cada grupo, la nueva unidad de análisis es una variable cuantitativa que puede tomar valores continuos entre 0 y 1. El cambio de unidad de análisis pasó del nivel individual a los grupos quinquenales. Este tipo de transformaciones son naturales en el análisis de la información. Nuestro objeto “ggtasas” es un objeto de tipo *dataframe*, que reúne valores resumidos según grupos quinquenales y sexo. Cada línea representa un grupo de edad para cada categoría de sexo.

## F. Análisis descriptivo. La utilidad de los datos ponderados

A menos que nuestra base de datos sea fruto de un muestreo completamente aleatorio, sin estratos ni conglomerados, el análisis de la muestra tendría un comportamiento idéntico al de la población y sólo cambiaría el volumen. Pero, como ya se señaló en el capítulo II, el muestreo de la ENOE tiene un diseño complejo. Parte de este diseño se puede observar en el factor de expansión.

El factor de expansión es un valor numérico por el cual se debe multiplicar cada uno de los casos de una encuesta para obtener el total de la población.

El comando “tally()”, de la paquetería *dplyr*, nos permite hacer conteos rápidamente. Observa en los siguientes comandos el conteo de casos por fila —sin poner ninguna variable— y el conteo según el factor de expansión.

```
# Conteo de casos
base_descriptiva %>%
  tally()

##          n
## 1 397600

# Conteo de La población
base_descriptiva %>%
  tally(fac)

##          n
## 1 126078860
```

Es decir, 397 600 entrevistas están representando a 126 078 860 personas que conforman la población estimada para el tercer trimestre de 2019.

### a. Replicando los datos del INEGI

Cuando revisamos los valores en los tabulados oficiales de las encuestas, normalmente observamos que los valores son mucho más grandes que con los que hemos trabajado. Hasta este momento hemos utilizado la información de los entrevistados, pero cada entrevistado fue seleccionado para representar a la población total del país.

Para obtener los valores poblacionales, utilizamos el factor de expansión, que “expande” cada una de nuestras observaciones. En términos probabilísticos, el factor de expansión se calcula como el inverso de la probabilidad de selección e indica el valor que representa el elemento seleccionado en la población total.

Tomando en cuenta lo anterior, vamos a replicar algunos tabulados del INEGI para los totales, tal como se publican en su sitio <<https://www.inegi.org.mx/programas/enoe/15ymas/default.html#-Tabulados>>.

En específico, retomaremos los indicadores calculados para la población de 15 años y más según tipo de inserción. Es importante usar los filtros desarrollados por el INEGI. Las variables precodificadas de “clase” ya retoman información del INEGI, por lo que sí haremos tabulados con algunos de ellas. Es importante eliminar los “no aplica” con un filtro de edad, puesto que estas variables asumen que se trata de personas en edad para trabajar, definidas como de 15 años y más y excluyendo a quienes no tienen edad definida (edad==99).

El comando “tabyl()”, del paquete *janitor*, es muy útil pero no es compatible con los factores de expansión. En realidad, con “tabyl()” ahorramos un poco de tiempo al agrupar nuestra base en categorías, aunque luego hagamos el conteo de cada una de ellas. “tally()” es un comando que hace ese conteo y “group\_by()” agrupa las observaciones de nuestra base de datos para hacer cualquier operación.

```
# Conteo de casos
base_descriptiva %>%
  filter(edad>14 & edad<99) %>% # Este filtro es muy importante
  mutate(clase1=as_label(clase1)) %>% # para usar etiquetas
  group_by(clase1) %>% # agrupa la base
  tally() # da conteos para grupos

## # A tibble: 2 x 2
##   clase1     n
##   <fct>    <int>
## 1 PEA     185002
## 2 PNEA    116342
```

La ventaja de “tally()” es que podemos ponerle un peso a su interior y en lugar de contar casos puede sumar variables, tal como sucede con el factor de expansión:

```
# Conteo de factor de expansión
base_descriptiva %>%
  filter(edad>14 & edad<99) %>% # Este filtro es muy importante
  mutate(clase1=as_label(clase1)) %>% # para usar etiquetas
  group_by(clase1) %>% # agrupa la base
  tally(fac) # suma el factor de expansión

## # A tibble: 2 x 2
##   clase1     n
##   <fct>    <dbl>
## 1 PEA      57349577
## 2 PNEA     37597058
```

Estos valores ya expandidos coinciden con los tabulados del INEGI. Es una buena práctica que revisemos siempre los resultados contra los tabulados publicados para saber si estamos calculando los elementos correctamente o si nuestras diferencias se basan en algún cambio de criterio.

Algunas opciones de *janitor* se pueden agregar para obtener los totales como una nueva fila:

```
# Conteo de factor de expansión
base_descriptiva %>%
  filter(edad>14 & edad<99) %>% # Este filtro es muy importante
  mutate(clase1=as_label(clase1)) %>% # para usar etiquetas
  group_by(clase1) %>% # agrupa la base
  tally(fac) %>% # suma el factor de expansión
  adorn_totals("row") # agrega una fila con totales

## #> #> clase1     n
## #> #>   PEA 57349577
## #> #>   PNEA 37597058
## #> #>   Total 94946635
```

Finalmente, calculamos las proporciones con “adorn\_percentages()” y agregamos el formato de “%”.

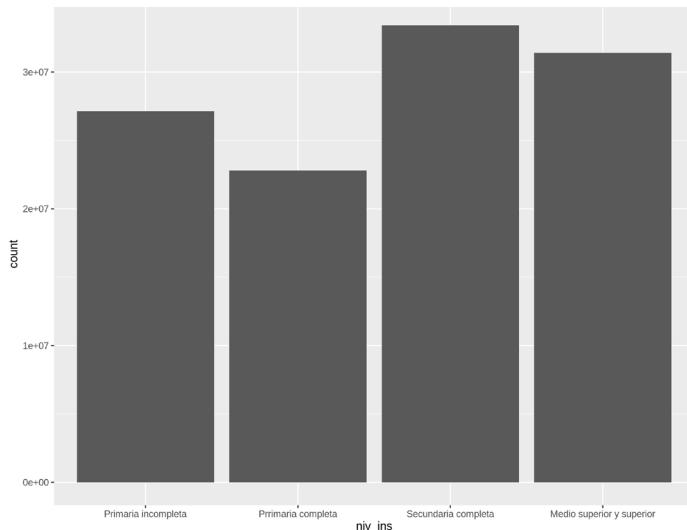
```
base_descriptiva %>%
  filter(eda>14 & eda<99)%>%
  mutate(clase1=as_label(clase1)) %>%
  group_by(clase1) %>%
  tally(fac) %>%
  adorn_totals("row") %>%
  adorn_percentages("all") %>%
  adorn_pct_formatting()

## #> #> clase1      n
## #>   PEA    60.4%
## #>   PNEA   39.6%
## #>   Total  100.0%
```

¿Qué pasa con nuestras gráficas? Para agregar el factor a las gráficas, colocamos “pesos” en ellos. Tomemos nuestra gráfica más sencilla de escolaridad: agregamos una opción dentro de “aes()” de “weight=fac”.

```
g + geom_bar(aes(weight=fac))
```

**Gráfica III-21. Barras de los niveles de instrucción.  
Conteo de datos expandidos**



FUENTE: elaboración propia con datos de la ENOE.

Para utilizar todo el diseño muestral de varias etapas y conglomerados, no sólo debemos ajustar por el factor de expansión nuestras estimaciones, por lo que más adelante haremos uso de la paquetería que nos permite introducir estos elementos en nuestras estimaciones: *srvyr*.

En este capítulo, hicimos un breve recorrido por el análisis descriptivo ad hoc a las variables cualitativas. En este proceso también revelamos algunas desigualdades por género en los mercados laborales mexicanos. En el siguiente capítulo seguiremos analizando estas disparidades, pero a partir de una variable cuantitativa muy particular: los ingresos laborales.

## **IV. El análisis descriptivo de las variables numéricas. El caso de los ingresos laborales en la ENOE**

### **Introducción**

Sin duda alguna, en un sistema regido por el mercado, necesitamos ingresos monetarios para comprar bienes y servicios. Como ya se mencionó en la introducción, los ingresos en México provienen fundamentalmente del trabajo. Los ingresos por trabajo (subordinado o independiente) representan el 78.65% de los ingresos monetarios de los hogares y el 64.90% de los ingresos corrientes que también consideran los flujos no monetarios de bienes y servicios (INEGI, 2018). Otras fuentes de ingreso como las rentas, las que provienen de intereses, o las transferencias gubernamentales o privadas no son tan importantes como en otras sociedades (Medina & Galván, 2008).

Por ello, parece importante ejemplificar el análisis descriptivo con esta variable. De ella parten muchas desigualdades y disparidades. Esta variable es numérica y continua. Su escala indica que el aumento en una unidad se mantiene a lo largo de su distribución; es decir, tenemos muy claro que un peso es un peso. Por ejemplo, en las escalas ordinales esto no queda tan claro; cuando pasábamos de secundaria completa a nivel superior, había cambios diferentes que cuando pasábamos de ninguna a primaria incompleta.

Las variables numéricas, tanto de intervalo como de razón, nos permiten hacer una gran cantidad de operaciones, tal como lo vimos en el Cuadro III-1. El análisis descriptivo de los datos cuantitativos implica hacer cálculos más sofisticados que las tablas de frecuencias que vimos anteriormente. Por ello, en lugar de comenzar con las operaciones, en este capítulo primero veremos algunas gráficas, iniciando por el análisis descriptivo de una sola variable en la primera sección. En una segunda sección, nos concentraremos en la construcción de la variable de ingresos laborales y algunos problemas

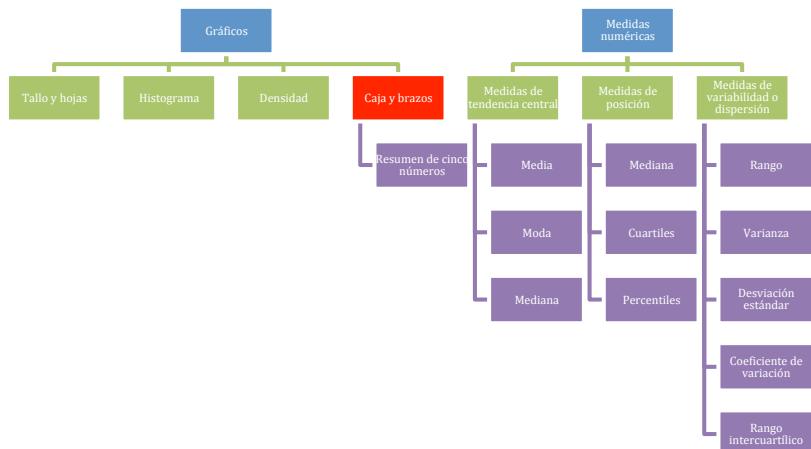
que tiene su captación. La tercera sección hace una introducción al análisis bivariado al ejemplificar cómo compararíamos una variable cuantitativa con otras categóricas.

En una cuarta sección revisamos cómo se calcula el índice de Gini, una medida muy utilizada para dar cuenta de la distribución de los ingresos; ello debido a que la desigualdad es una palabra que fue sumamente relevante en los estudios analizados en el capítulo I. Finalmente, en la última sección, revisamos la relación entre dos variables cuantitativas que sienta una muy buena introducción para el capítulo VI, donde revisamos más sobre esta relación con el caso de la regresión lineal simple.

## A. Análisis descriptivo univariado

En esta sección haremos un recorrido no exhaustivo pero que sí intenta concentrarse en las medidas y gráficas más comunes. La Figura IV-1 muestra un listado que guiará la exposición. Iniciaremos con algunas gráficas, a excepción de la de caja y brazos que necesita primero comprender algunas medidas numéricas para su construcción.

**Figura IV-1. Esquema de contenido de técnicas descriptivas univariadas clásicas a discutir**



FUENTE: elaboración propia.

De nuevo, la recomendación es revisar el sitio <<https://www.data-to-viz.com/>> para un menú más amplio de gráficas. Además, recomendamos la lectura de libros de estadística para complementar los detalles que se mencionan. Dos recomendaciones completamente subjetivas de quien escribe este libro son las obras de Moore (Moore, 2010) y de Mendenhall, Beaver y Beaver (2014); estos textos se caracterizan por su alto contenido didáctico y una gran cantidad de ejemplos detallados.

Antes de iniciar con las gráficas, vamos a cargar los paquetes, o instalarlos si es necesario. Cargamos la base que obtuvimos en el capítulo anterior.

```
if (!require("pacman")) install.packages("pacman") # instala pacman si se requiere
## Loading required package: pacman

#cargan los paquetes necesarios para la práctica de este capítulo
pacman::p_load(tidyverse, haven,
  janitor,
  svyr,
  RColorBrewer, wesanderson,
  sjlabelled,
  ineq, gglorenz)
load("completat319.RData") ## Recuerda tenerla descargada en tu carpeta de trabajo
```

Para hacer más sencilla esta práctica, utilizaremos de nuevo un subconjunto de nuestra base de datos que selecciona algunas variables:

```
base_descriptiva<-completat319 %>%
  select(c(sex, eda, niv_ins, t_loc, anios_esc, #vars socdem
    clase1, clase2, clase3, # inserción Laboral
    rama, c_ocu11c, pos_ocu, hrsocup, imssissste, #inserción
    ing7c, ingocup, ing_x_hrs, # ingresos
    p6b1, p6b2, p6c, # vars coe - control
    fac, est_d, upm)) #diseño muestral

rm(completat319)
```

## a. Las primeras gráficas

A continuación presentamos tres tipos de gráficas muy básicas que se aconseja revisar para cualquier variable cuantitativa. Estos tipos de gráficas nos permiten establecer:

- la dispersión y rango de la variable;
- la asimetría (derecha o positiva; o bien, izquierda o negativa);

- c) dónde se concentran los datos y por tanto una aproximación a su centro; y, finalmente,
  - d) la presencia de *outliers* o datos atípicos.

Estos son cuatro elementos básicos que leemos en la distribución de valores de cualquier variable cuantitativa y son los elementos mínimos que se deben de retomar para su análisis.

Como son más elementos que en las gráficas cualitativas, recomendamos tomar en cuenta el nivel de análisis; es decir, la unidad en que se registró la medición (individuos, estados, países, sectores) y las unidades de medida de la variable (pesos, años cumplido, etcétera).

i) Gráfica de tallo y hoja

La gráfica de tallo y hoja es una representación de un “tallos”, establecido por decenas o algún múltiplo de ellas, y “hojas” que representan cada uno de los datos en análisis. Esto nos permite dar una mirada rápida a elementos clave de la distribución de la variable en estudio. Haremos una gráfica de la variable “ingresos por hora”, que ya viene computada por el INEGI, para los ocupados (“clase2==1”) y quienes hayan reportado un ingreso mayor a 0 (“ing\_x hrs>0”):

Esta gráfica señala que las decenas se han hecho a dos dígitos, es decir, que el “42” que observamos al final de la gráfica corresponde a quienes ganan poco más de 4 200 pesos la hora trabajada.

```
max(base_descriptiva$ing_x_hrs) # nos da el valor máximo
## [1] 4375
4375 - 4200
## [1] 175
175 / 20 # de aquí proviene el 8
## [1] 8.75
```

Claramente, este caso no es similar a la mayoría. Datos que salen tanto de la norma del resto se llaman *outliers* o datos atípicos.

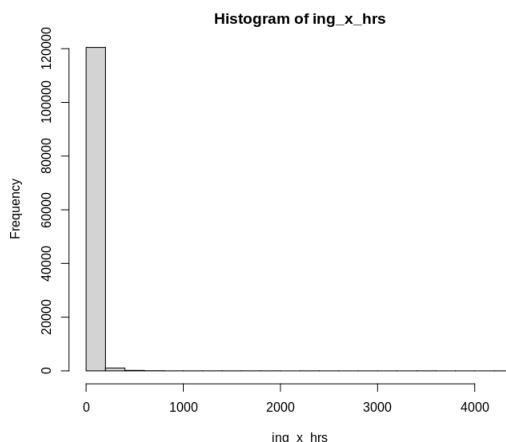
### *ii) Histograma*

Un histograma es una manifestación gráfica de barras que representan intervalos de una variable cuantitativa y sus frecuencias, frecuencias relativas o densidad. Las barras no están separadas puesto que representan intervalos que se unen en sus límites y, por lo general, son del mismo color.

Para realizarlo, usamos los siguientes códigos:

```
base_descriptiva %>%
  filter(clase2==1 & ing_x_hrs>0) %>%
  with(hist(ing_x_hrs))
```

**Gráfica iv-1. Histograma de los ingresos por hora, calculado con la función de base. Trimestre III, 2019. Pesos por hora**

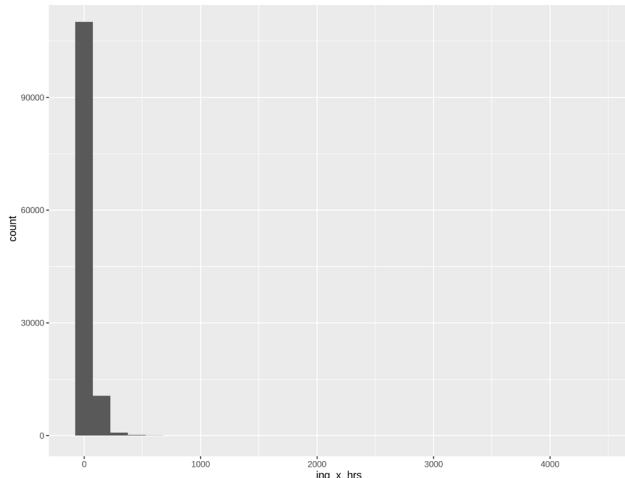


FUENTE: elaboración propia con datos de la ENOE.

En “ggplot()”, el histograma se realiza con los siguientes códigos:

```
base_descriptiva %>%
  filter(clase2==1 & ing_x_hrs>0) %>%
  ggplot(aes(ing_x_hrs)) + geom_histogram()
```

**Gráfica iv-2. Histograma de los ingresos por hora, calculado con “ggplot()”. Trimestre III, 2019. Pesos por hora**



FUENTE: elaboración propia con datos de la ENOE.

En este tipo de gráfica se ve rápidamente el rango, marcado por lo ancho de la gráfica, y la moda o el dato que más repite, que es donde tenemos mayor acumulación de frecuencias. Por otro lado, volvemos a observar el sesgo a la derecha ocasionado por un grupo pequeño de observaciones que tienen muy altos ingresos. Lo cual reflejaría sin duda la desigualdad dentro del mercado laboral mexicano. Además, definitivamente tenemos *outliers* o valores atípicos que ganan más de 4 000 pesos por hora.

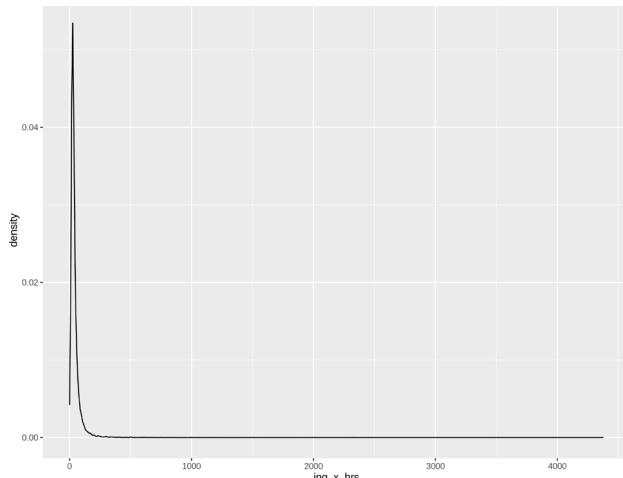
### *iii) Gráfica de densidad*

Una gráfica muy popular es la de densidad. La densidad es una medida de probabilidad a la que se llega a partir de una aproximación de

la frecuencia relativa. La opción “geom\_density”, de acuerdo con la ayuda de *R*, señala que “calcula y dibuja la estimación de la densidad Kernel, que es una versión suavizada del histograma”. Esta es una alternativa útil al histograma para datos continuos que provienen de una distribución uniforme subyacente.

```
base_descriptiva %>%
  filter(clase2==1 & ing_x_hrs>0) %>%
  ggplot(aes(ing_x_hrs)) + geom_density()
```

**Gráfica iv-3. Línea de densidad de los ingresos por hora, calculado con la función de base. Trimestre III, 2019. Pesos por hora**

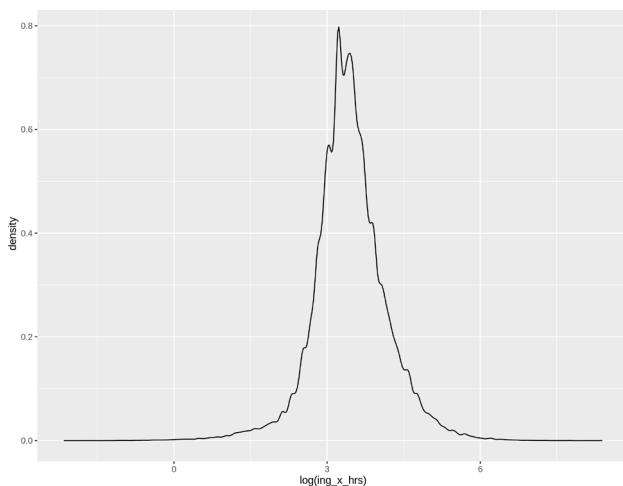


FUENTE: elaboración propia con datos de la ENOE.

La ventaja de *R* es que podemos introducir operaciones en nuestra variable. A veces, cambiar la escala de nuestra variable nos permite visualizar elementos fundamentales en el análisis estadístico. Una transformación tiene que ver con la forma; por ejemplo, calcular el logaritmo de la variable:

```
base_descriptiva %>%
  filter(clase2==1 & ing_x_hrs>0) %>%
  ggplot(aes(log(ing_x_hrs))) +
  geom_density()
```

**Gráfica iv-4. Línea de densidad de los ingresos por hora, calculado con la función de base. Trimestre III, 2019. Escala logarítmica**



FUENTE: elaboración propia con datos de la ENOE.

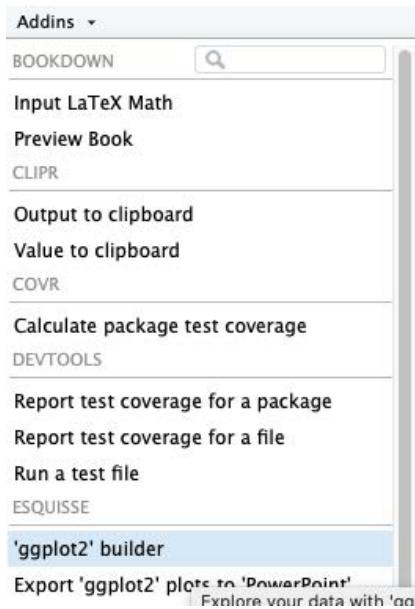
### b. Un atajo *esquisse*

Para usar este complemento, lo mejor es generar un subconjunto con pocas variables para que sea más fácil de leer desde la herramienta. Para ello, haremos una subbase ya filtrada:

```
b_esquisse<-base_descriptiva %>%
  filter(clase2==1 & ing_x_hrs>0) %>% #ingresos válidos y mayores que 0
  mutate(sex=as_label(sex)) %>% # variable con las etiquetas
  select(ing_x_hrs, sex, eda) # Algunas variables interesantes
```

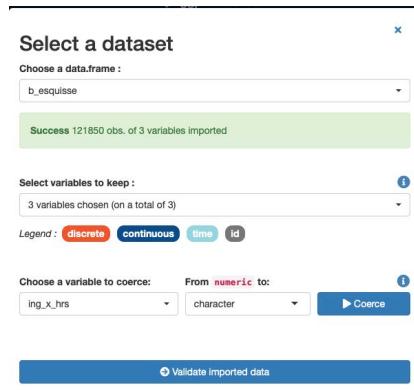
Si la lógica de las capas y lienzos parece un poco complicada para graficar con “`ggplot()`”, se puede utilizar el paquete *esquisse*, el cual es un programa que despliega un menú interactivo mucho más amigable con el usuario. Una vez que se instala se accede a él en la sección de complementos, ubicada en la barra de tareas alineada con los botones de guardar y de búsqueda; ahí damos un clic en “Add ins” (complemento, en español) y se despliega la siguiente imagen:

**Figura iv-2. Menú de RStudio donde se muestran algunos complementos que pueden instalarse, entre ellos el esquisse**



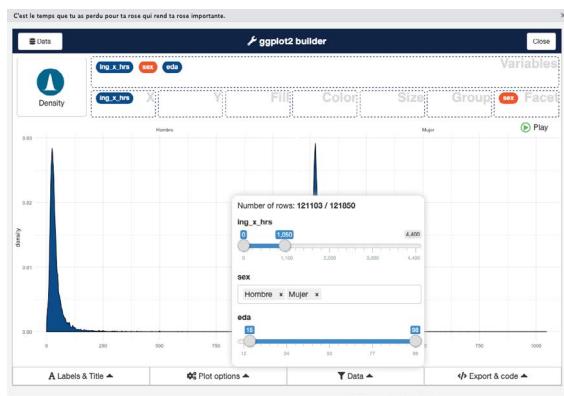
Desde el complemento se accede al siguiente menú para seleccionar nuestra base aún más pequeña.

**Figura iv-3. Menú de selección de datos de esquisse**



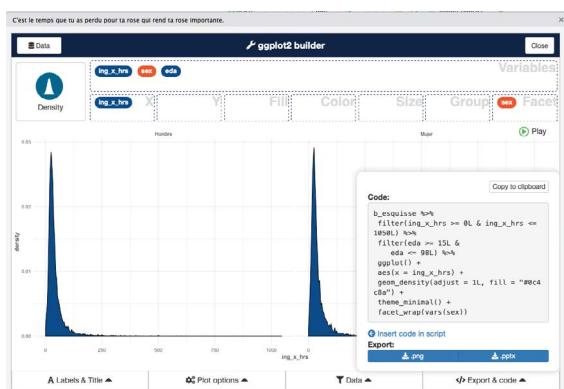
Una vez seleccionada, se hace un “drag and drop” (arrastrar y soltar) de las variables y se elige el tipo de gráfica, incluso permite hacer un filtro con las variables disponibles.

**Figura iv-4. Módulo interactivo del paquete**



Esto se puede exportar a una imagen, a Power Point o como código para ejecutar en *RStudio*. Esta herramienta es muy útil para acostumbrarnos al código compatible con *ggplot2*, que puede ser bastante complicado en los primeros acercamientos.

**Figura iv-5. Ejemplo de cómo se puede exportar el código de la herramienta interactiva**



## c. Las medidas numéricas: la media y la desviación estándar

Una de las medidas más comunes para establecer el centro de la distribución es el promedio o media aritmética: la suma de todos los valores de nuestra variable dividida entre el total de observaciones. La media tiene varias propiedades, por ejemplo, si sumamos todas las desviaciones a este valor, la suma de ellas es cero.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

Donde  $n$  es el tamaño de la muestra, el subíndice  $i$  representa cada uno de los individuos y  $x$  la medición de nuestra variable. En el caso de la media poblacional,  $\mu$ ,  $N$  es el tamaño de la población.

Para su cálculo utilizamos la función “mean()”, de base, pero con el comando “summarise” de *dplyr* para obtener el resultado de una operación después de otra, a través de los *pipes*:

```
base_descriptiva %>%
  filter(clase2==1 & ing_x_hrs >0) %>% #filtro de casos
  summarise(promedio=mean(ing_x_hrs)) #hace un "resumen" con la media
##   promedio
## 1 41.46826
```

El promedio es una medida de centro, pero el centro nos dice poco si no sabemos cómo se comportan las observaciones alrededor de él. Para ello, necesitamos medidas de dispersión y su mejor acompañante es la desviación estándar.

Antes de calcular la desviación estándar, debemos calcular la varianza. Para ello, de nuevo necesitamos el concepto de desviación, que es la diferencia de un valor con respecto a una norma. Por lo general, asumimos esta norma como la media aritmética. Del mismo modo, el promedio aritmético de los cuadrados de las desviaciones de los valores de la variable con respecto a una constante cualquiera se hace mínima cuando dicha constante coincide con la media aritmética.

La varianza nos da una medida de distancia promedio sin el problema que siempre dé cero, como pasaría si no se eleva al cuadrado. A continuación, presentamos sus fórmulas para la población y la muestra.

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

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{N}$$

La medida muestral es diferente a la poblacional en su denominador. Esto proviene de la corrección de Bessel, que corrige el sesgo estadístico en la estimación de la varianza poblacional.

La varianza es una medida muy importante, pero difícil de interpretar debido a que tenemos las unidades originales de nuestra variable al cuadrado: pesos al cuadrado, años al cuadrado, horas al cuadrado. De ahí que sea importante sacarle raíz cuadrada:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

La desviación estándar es, entonces, una medida de dispersión que nos dice qué tan alejados están los datos de la media, por lo que aporta mucha más información que la media sola. Por lo general, las colocamos juntas. Podemos pedir un resumen con *dplyr* para que nos muestre una tabla con las medidas juntas:

```
base_descriptiva %>%
  filter(clase2==1 & ing_x_hrs >0) %>% #filtro de casos
  summarise(media=mean(ing_x_hrs), # media
            sd=sd(ing_x_hrs), # desviación estándar
            var=var(ing_x_hrs)) #varianza

##      media      sd      var
## 1 41.46826 50.70867 2571.37
```

Estos resultados nos dicen que, en promedio, los entrevistados que reportaron ingresos mayores a cero ganan 41.46 pesos la hora y que tienen una desviación estándar de 50.70 pesos. El solo hecho de

que la desviación supere la media ya nos está diciendo que esta variable es sumamente heterogénea. La varianza es de 2 571.37 pesos al cuadrado, lo cual es un valor difícil de imaginar.

Las funciones que acabamos de revisar brindan las estimaciones muestrales. Si necesitas las estimaciones poblacionales, aplicaremos el artilugio de multiplicar por  $(N - 1/N)$  para que se elimine el denominador  $(N - 1)$  y quede multiplicado por  $N$ . Esto se puede hacer con la función “length()” para nuestro vector de análisis “ing\_x\_hrs”.

Asumiendo que tuviéramos una población y no una muestra.

```
base_descriptiva %>%
  filter(clase2==1 & ing_x_hrs >0) %>% #filtro de casos
  summarise(media=mean(ing_x_hrs), # media
            var.p=var(ing_x_hrs)*(length(ing_x_hrs)-1)/length(ing_x_hrs), #varianza
            sd.p=sqrt(var(ing_x_hrs)*(length(ing_x_hrs)-1)/length(ing_x_hrs))) # desviación
estándar
##      media    var.p     sd.p
## 1 41.46826 2571.349 50.70847
```

Las diferencias entre las estimaciones poblacionales y las muestrales son muy pocas porque la muestra es grande y las diferencias en el denominador generan pocos cambios.

#### d. El resumen de cinco números y las gráficas de caja y brazos

La media es una medida muy popular, pero tiene el problema de que está afectada por los valores atípicos. Como hemos observado en las gráficas realizadas en el acápite a de esta sección, este es un problema en la distribución de los ingresos por trabajo en México.

Por ejemplo, la media para la totalidad de los datos es:

```
base_descriptiva %>%
  filter(clase2==1 & ing_x_hrs >0) %>% #filtro de casos
  summarise(media=mean(ing_x_hrs))

##      media
## 1 41.46826
```

Pero lo que observamos visualmente en las gráficas es a partir de 1 000 pesos, los datos se alejan mucho de la norma. Calculemos sin esos valores y calculemos cuántos casos son:

```
base_descriptiva %>%
  filter(clase2==1 & ing_x_hrs >0 & ing_x_hrs<1000) %>% #filtro de casos y quitando
atípicos del Lado derecho
  summarise(media=mean(ing_x_hrs))

##      media
## 1 41.08117

base_descriptiva %>%
  filter(clase2==1 & ing_x_hrs >0) %>% #filtro de casos
  count(ing_x_hrs>999) # Conteo de atípicos

## # A tibble: 2 x 2
##   `ing_x_hrs > 999`     n
##   <lgl>                <int>
## 1 FALSE                 121819
## 2 TRUE                  31
```

Vemos cómo 31 casos de una muestra de 121 850 causan un aumento de casi 40 centavos por hora en el promedio.

Otra medida popular y más robusta a los *outliers* es la mediana. Ella representa el valor de la variable en la posición central en un conjunto de datos ordenados. Es decir, supera al 50% de los casos y su valor es superado en el 50% restante. Es, por tanto, una medida de centro y una medida de posición.

Veamos qué pasa si en lugar de la media usáramos la mediana para establecer el centro:

```
base_descriptiva %>%
  filter(clase2==1 & ing_x_hrs >0) %>% #filtro de casos
  summarise(mediana=median(ing_x_hrs))

##      mediana
## 1 30.14643

base_descriptiva %>%
  filter(clase2==1 & ing_x_hrs >0 & ing_x_hrs<1000) %>% #filtro de casos y quitando
atípicos del Lado derecho
  summarise(mediana=median(ing_x_hrs))

##      mediana
## 1 30.14643
```

La mediana no cambia en las dos situaciones, con filtro y sin filtro de valores atípicos (y tiene un valor mucho menor que nuestra media). De ahí que muchas veces se prefiera este estimador a la media, puesto que está menos afectada por los valores atípicos. Pero, además, esta cualidad también la convierte en un mejor estimador del centro de los datos cuando tenemos una distribución sesgada.

Que la media supere a la mediana da información sobre el sesgo a la derecha que mantiene la distribución. Si los valores son iguales o muy cercanos, seguro estamos ante una distribución bastante simétrica; mientras que, si la mediana supera a la media, ello da cuenta de que existen valores a la izquierda de la distribución que la están sesgando, de ahí que podemos aducir la existencia de un sesgo negativo.

Cuando tenemos esta situación, la media no es tan representativa y, para comprender más nuestra distribución, necesitamos medidas que acompañen a una medida de centro como la mediana. De ahí proviene la necesidad del resumen de cinco números:

```
base_descriptiva %>%
  filter(clase2==1 & ing_x_hrs >0 ) %>% #filtro de casos y quitando atípicos del Lado
derecho
  with(summary(ing_x_hrs))

##      Min.    1st Qu.     Median      Mean    3rd Qu.      Max.
##  0.116   20.930   30.146   41.468   46.242  4375.000
```

Con el comando “summary()” obtenemos estos seis números (se agrega la media). El resto es lo que conocemos como el resumen de cinco números y se incluye el mínimo y máximo en los extremos y otras dos medidas de posición, además de la mediana, el cuartil 1 ( $Q_1$ ) y el cuartil 3 ( $Q_3$ ). El cuartil 1 es una medida de posición, igual que la mediana, que separa la población en un 25% inferior y un 75% superior; mientras que el cuartil 3 separa a la población en un 75% inferior y un 25% superior.<sup>13</sup> Estas medidas dan una idea de cómo se distribuye nuestra variable, pero también son la base de una de los gráficas más famosas: la gráfica de caja y brazos, o *box plot*.

La gráfica de caja y brazos (o caja y bigotes) también toma en cuenta el concepto de rango intercuartílico (*RIC*), que es la diferencia entre el cuartil 1 y el cuartil 3; es decir, establece el rango donde se concentra el 50% de los datos.

$$RIC = Q_3 - Q_1$$

---

<sup>13</sup> En este sentido, la mediana es el cuartil 2 ( $Q_2$ ).

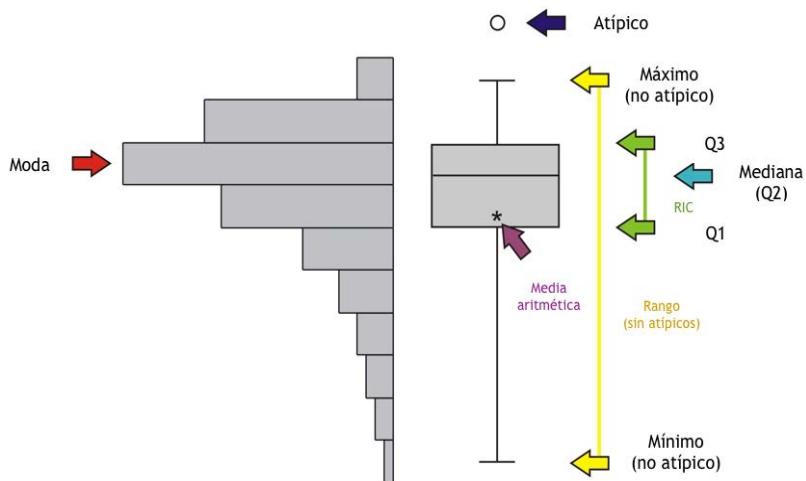
Otro concepto que utiliza esta gráfica es el de atípicos, pero más allá de lo que ya habíamos hecho antes (visualmente). Propone unas medidas de límites inferior y superior:

$$L_{inferior} = Q_1 - 1.5 * RIC$$

$$L_{superior} = Q_3 + 1.5 * RIC$$

Es decir, cualquier dato será atípico si es menor a  $L_{inferior}$  y mayor a  $L_{superior}$ . De tal forma, podemos resumir todo esto y compararlo con el histograma en la siguiente imagen:

**Figura iv-6. Comparación de un histograma con una gráfica de caja y brazos**

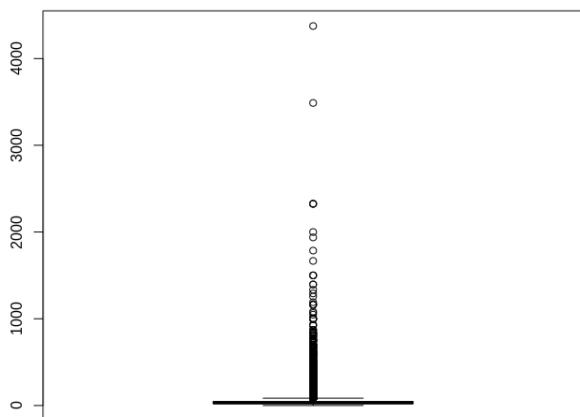


FUENTE: Wikipedia (2020).

Para graficarlo en R, tenemos el siguiente código que se vale de la función “boxplot()”:

```
base_descriptiva %>%
  filter(clase2==1 & ing_x_hrs >0 ) %>% #filtro de casos y quitando atípicos del Lado
derecho
  with(boxplot(ing_x_hrs))
```

**Gráfica iv-5. Diagrama de caja y brazos de los ingresos por hora, calculado con la función de base. Trimestre III, 2019.**



FUENTE: elaboración propia con datos de la ENOE.

Tenemos tantos casos atípicos y un sesgo positivo tan grande que casi no podemos verlo. Esta gráfica también se puede realizar desde *esquisse*. Es algo sencillo y verifica que se entiende el código para “`ggplot()`”.

Definitivamente, hay más medidas para las variables cuantitativas, pero hemos reseñado un grupo básico para su análisis. A continuación revisaremos más la variable de los ingresos laborales de la ENOE y su cálculo.

## B. Los ingresos laborales en la ENOE

Las variables ya calculadas por el INEGI que refieren a los ingresos de los trabajadores (“`ing_x_hrs`”, “`ingocup`”) provienen de una serie de operaciones resultantes de preguntas específicas en el cuestionario básico.

Este proceso queda clarificado siguiendo lo que se establece en el documento “Encuesta Nacional de Ocupación y Empleo. Reconstrucción de variables. 2005 a la fecha” (INEGI, 2015). Ahí aprendemos que, dependiendo de la frecuencia del pago o de cómo la persona considera sus ingresos, se tendrán que hacer operaciones para llegar a los estimados mensuales.

Todo ello se ve claramente en la Figura IV-7, en la que mostramos las preguntas necesarias para llegar a los ingresos. Hay varias problemáticas que deben tomarse en cuenta a la hora de estudiar los ingresos. Entre ellas, la gente que no reporta ingresos porque en realidad participa en la actividad económica sin remuneración. Pero aún más importante, existe un grupo de personas a lo largo de la historia de la encuesta que no ha contestado cuánto gana de manera numérica. Algunos contestan con intervalos por tramos en términos de salarios mínimos y otros no han contestado nada acerca de los ingresos. Éste es un fenómeno que ha ido en aumento a lo largo del tiempo. La figura muestra recuadros de algunos elementos asociados a estos fenómenos, los que describiremos paso a paso.

**Figura IV-7. Preguntas clave para identificar valores perdidos de ingresos. Página 8 del cuestionario básico**

<p><b>5f. ¿En qué meses del año ... realiza este trabajo?</b>  <small>(Escucha y circula, según la respuesta del informante)</small></p> <table style="width: 100%; border-collapse: collapse;"> <tr><td style="width: 33.33%;">01 Enero</td><td style="width: 33.33%;">02 Febrero</td><td style="width: 33.33%;">03 Marzo</td></tr> <tr><td>04 Abril</td><td>05 Mayo</td><td>06 Junio</td></tr> <tr><td>07 Julio</td><td>08 Agosto</td><td>09 Septiembre</td></tr> <tr><td>10 Octubre</td><td>11 Noviembre</td><td>12 Diciembre</td></tr> </table> <p>13 Varian los meses en que trabaja      14 Trabaja todos los meses del año      15 Tiene menos de un año en este trabajo      99 NS  <b>00 Exclusivo capturista</b></p>	01 Enero	02 Febrero	03 Marzo	04 Abril	05 Mayo	06 Junio	07 Julio	08 Agosto	09 Septiembre	10 Octubre	11 Noviembre	12 Diciembre	<p><b>6b. ¿Cada cuándo obtiene ... sus ingresos o le pagan?</b>  <small>(Escucha, clasifica el periodo, pregunta por los ingresos y anótalo)</small></p> <p><b>¿Cuánto ganó o en cuánto calcula sus ingresos?</b></p> <table style="width: 100%; border-collapse: collapse;"> <tr><td style="width: 66.66%;">1 Cada mes</td><td style="width: 33.33%;">\$ _____</td></tr> <tr><td>2 Cada 15 días</td><td>\$ _____</td></tr> <tr><td>3 Cada semana</td><td>\$ _____</td></tr> <tr><td>4 Diario</td><td>\$ _____</td></tr> <tr><td>5 Otro periodo de pago</td><td>\$ _____</td></tr> <tr><td>_____ Periodo</td><td>Pasa a la d</td></tr> </table> <p>6 Le pagan por pieza producida o vendida, servicio u obra realizada</p> <table style="width: 100%; border-collapse: collapse;"> <tr><td style="width: 66.66%;">Unitad</td><td style="width: 33.33%;">Precio por unidad</td></tr> <tr><td colspan="2">Total de unidades por semana _____</td></tr> <tr><td colspan="2">_____</td></tr> <tr><td colspan="2">_____</td></tr> </table>	1 Cada mes	\$ _____	2 Cada 15 días	\$ _____	3 Cada semana	\$ _____	4 Diario	\$ _____	5 Otro periodo de pago	\$ _____	_____ Periodo	Pasa a la d	Unitad	Precio por unidad	Total de unidades por semana _____		_____		_____	
01 Enero	02 Febrero	03 Marzo																															
04 Abril	05 Mayo	06 Junio																															
07 Julio	08 Agosto	09 Septiembre																															
10 Octubre	11 Noviembre	12 Diciembre																															
1 Cada mes	\$ _____																																
2 Cada 15 días	\$ _____																																
3 Cada semana	\$ _____																																
4 Diario	\$ _____																																
5 Otro periodo de pago	\$ _____																																
_____ Periodo	Pasa a la d																																
Unitad	Precio por unidad																																
Total de unidades por semana _____																																	
_____																																	
_____																																	
<p><b>VI. INGRESOS Y ATENCIÓN MÉDICA</b></p> <p><b>6. ¿ ... recibe o le pagan</b>  <small>(Lee las opciones y circula las indicadas por el informante)</small></p> <p>01 por comisión?      02 a destajo (por pieza), servicio u obra realizada?      03 por honorarios?      04 con propinas?      05 con bonos de compensación o de productividad?      06 con vales o productos comerciales?      07 Solo recibe sueldo, salario o jornal      08 Solo lo que le deja su negocio → Pasa a 6d      09 No le pagan ni recibe ingresos → Pasa a 6d  <small>(incluye autoconsumo agropecuario)</small>      10 Ninguna de las anteriores      99 NS  <b>00 Exclusivo capturista</b></p>																																	
<p><b>6a. Aparte de lo que me acaba de mencionar, ¿ ... obtiene o le pagan sus ingresos</b>  <small>(Lee las opciones y circula las indicadas por el informante)</small></p> <p>1 a sueldo, salario o jornal?      2 por ganancias o de lo que deja su negocio?  <b>3 No le pagan ni recibe ingresos</b> → Pasa a 6d  <small>(incluye autoconsumo agropecuario)</small>      4 Ninguna de las anteriores      9 NS  <b>00 Exclusivo capturista</b></p>																																	
<p><b>6d. Por parte de este trabajo ¿ ... tiene acceso a atención médica en</b>  <small>(Lee las opciones y circula la indicada por el informante)</small></p> <p>1 el Seguro Social (IMSS)?      2 el hospital o clínica naval, militar o de Pemex?      3 el ISSSTE?      4 el ISSSTE estatal (ISSSTELEON, ISSEMYM)?      5 otra institución médica?      Específica      6 No recibe atención médica      9 NS</p>																																	

FUENTE: INEGI (2019).

Como ya señalamos, hay quienes realmente obtienen cero pesos porque son personas que no reciben pago (“p6a==3”) y otra parte de personas se rehusa a contestar en las preguntas p6b y p6c; o bien, no puede estimar sus ingresos.

Vamos a revisar muy bien esta variable puesto que, como ya se mencionó, es clave para el estudio de la desigualdad en México. Iniciamos con un conteo de quiénes están ocupados y tienen ingresos mayores a cero y quiénes tienen ingresos igual a cero:

```
base_descriptiva %>%
  filter(clase2==1) %>%
  count(ingocup>0)

## # A tibble: 2 × 2
##   `ingocup > 0`     n
##   <lgl>           <int>
## 1 FALSE            52772
## 2 TRUE             126520
```

¿Cuántos de estos ceros son válidos y cuántos son valores perdidos? Como señalamos, hay personas que no reciben pago (es posible identificarlas con la pregunta p6a y puede ser complementada con otras). Esto queda muy claro en la variable construida por el INEGI de “posición en la ocupación”, llamada “pos\_ocu”. Hagamos un tabulado de ella:

```
base_descriptiva %>%
  filter(clase2==1) %>% # filtro ocupados
  mutate(pos_ocu=as_label(pos_ocu)) %>% # etiquetas de La variable
  tabyl(pos_ocu, show_missing_levels=F) %>% # tabla de frecuencias
  adorn_totals("row") #agrega filas de totales

##      pos_ocu          n    percent
## Trabajadores subordinados y remunerados 126550  0.70583183
##           Empleadores      8698  0.04851304
## Trabajadores por cuenta propia      36053  0.20108538
## Trabajadores sin pago        7991  0.04456975
## Total                      179292 1.00000000
```

Por tanto, parte de estos ceros sí son válidos; al menos tienen la opción 4 en la variable “pos\_ocu”. Pero, como vemos, son sólo 7 991 de los 52 722 que tenemos.

Veamos quiénes tienen cero: no son trabajadores sin pago y además trabajaron más de una hora a la semana.

```
base_descriptiva %>%
  filter(clase2==1) %>% # filtro de ocupados
  filter(pos_ocu!=4) %>% # filtro para eliminar trabajadores sin pago
  count(ingocup==0) %>%
    mutate(percent=n/sum(n)*100)

## # A tibble: 2 x 3
##   `ingocup == 0`     n percent
##   <lgl>           <int>  <dbl>
## 1 FALSE            126520   73.9
## 2 TRUE             44781    26.1
```

Revisemos bien estos ceros. Primero en términos de la pregunta p6b1, donde las opciones 7 y 8 refieren a “No supo estimar”, “Se negó a contestar esta pregunta”, ante una estimación exacta de sus ingresos:

```
base_descriptiva %>%
  filter(clase2==1) %>% # filtro de ocupados
  mutate(p6b1=as_label(p6b1)) %>% # cambio de etiqueta
  tabyl(p6b1)

##                                     p6b1      n
## No aplica                      0
## Cada mes                       9161
## Cada 15 días                    27225
## Cada semana                     64985
## Diario                          21557
## Otro periodo de pago            2952
## Le pagan por pieza producida o vendida, servicio u obra realizada 640
## No supo estimar                 33443
## Se negó a contestar esta pregunta 8978
##                                         <NA> 10351

##   percent valid_percent
## 0.000000000 0.000000000
## 0.051095420 0.054226032
## 0.151847266 0.161150934
## 0.362453428 0.384660917
## 0.120234032 0.127600760
## 0.016464761 0.017473556
## 0.0033569596 0.003788305
## 0.186528122 0.197956683
## 0.050074738 0.053142813
## 0.057732637      NA
```

Podemos verificar que, al usar la etiqueta, tenemos una variedad de formas de pago. Una buena cantidad de personas no supieron contestar (33 443) y otras se negaron a contestar (8 978).

Pero con esta información y con la de posición en el empleo podemos verificar quiénes tienen “0” como un valor válido. Revisemos la siguiente secuencia de comandos, donde el filtro “pos\_ocu!=4 & p6b1<7” elimina a quienes podrían tener cero. Deberíamos de no contar con nadie que tenga ingresos cero:

```
base_descriptiva %>%
  filter(clase2==1) %>% # filtro de ocupados
  filter(pos_ocu!=4 & p6b1<7) %>% # filtro anterior + si no declaran ingresos
  count(ingocup==0) %>%
  mutate(percent=n/sum(n))

## # A tibble: 1 x 3
##   `ingocup == 0`     n percent
##   <lg1>           <int>    <dbl>
## 1 FALSE            126520      1
```

Tomando en cuenta esto, podemos crear una categoría con “NA” cuando los ceros no sean realmente cero y dejar el cero para el valor válido de quiénes no son remunerados. Lo haremos con el comando “mutate”, creando una nueva variable “ingocup” y además con el comando “ifelse”:

```
base_descriptiva<-base_descriptiva %>%
  mutate(ingocup2=
    ifelse( # función ifelse
      pos_ocu!=4 & p6b1>6 & ingocup==0, #si cumple con esta condición
      NA, #toma este valor
      ingocup)) # si no, toma éste
```

Ya construimos una variable que tiene los valores perdidos que se distinguen de los ceros. Esto es importante porque el análisis de la no respuesta de los ingresos puede ser un problema de investigación en sí mismo.

Hay varios enfoques para tratar los valores perdidos. Uno puede ser simplemente no tomarlos en cuenta en el análisis (que es lo que se ha venido haciendo a lo largo del capítulo al usar el filtro “ing\_x\_hrs>0”). Este proceso se llama *listwise*, es decir, se quitan de la lista todos los valores de este grupo en nuestro análisis. También podemos sustituir los valores por la media, ya sea general o de un grupo.

Otra opción tiene que ver con la imputación, es decir, calcular un valor a partir de un modelo o condiciones lógicas. En el caso mexicano, podemos recuperar cierta información de los intervalos de salarios mínimos.

No obstante, a veces se requiere hacer suposiciones y condiciones para más de una variable. De ahí que se necesiten modelos de imputación. Para esta opción, se utilizan otras herramientas estadísticas y otra serie de condiciones. Específicamente, para el caso mexi-

cano, hay aportes para la ENOE (Luján, 2009; Rodríguez & López, 2015) y para la ENIGH (Campos, 2013).

Algunos paquetes para desarrollar la imputación y de manera múltiple son:

- *mi* (Gelman, Hill, Su, Yajima, Pittau *et al.*, 2015)
- *hot.deck* (Cranmer, Gill, Jackson, Murr & Armstrong, 2020)
- *Amelia* (Honaker, King & Blackwell, 2019)
- *Mice* (Buuren, Groothuis, Vink, Schouten, Robitzsch *et al.*, 2020)

La aplicación de estos elementos excede los objetivos de este libro, pero se mencionan los paquetes que pudieran ayudar a quien desee adentrarse en el método de imputación.

## C. Comparando ingresos entre grupos

Para comparar cómo se comporta una variable cuantitativa a lo largo de una categórica, podemos repetir los elementos del análisis univariado y comparar todos los elementos que ya señalamos. ¿Cómo se comportan las categorías en el centro? ¿En términos de dispersión? ¿Se comportan igual o no en la asimetría? ¿Alguna categoría tiene más datos atípicos? Éstas son algunas de las preguntas que podríamos contestar con este análisis.

Para revisar estas particularidades, vamos a utilizar la variable “pos\_ocu”, con el fin de ver cómo diferentes formas de relacionarse con la ocupación y las relaciones de subordinación crean diferentes distribuciones de ingreso. Podemos repetir algunas de las gráficas ya estudiadas y colocar en un mismo lienzo varias distribuciones.

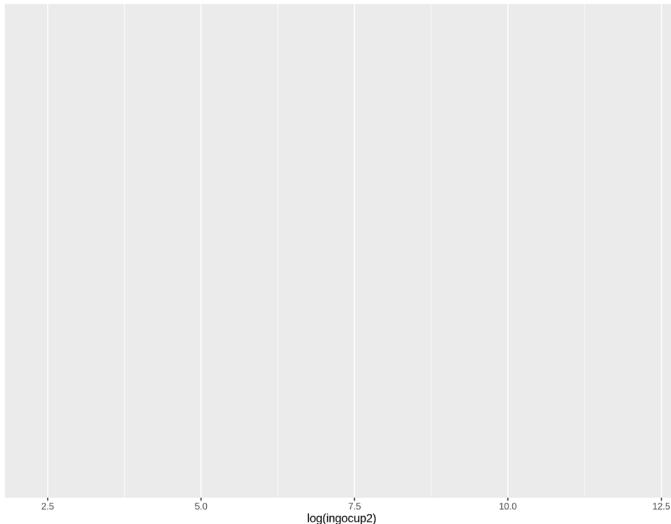
### a. Gráficas para mostrar diferentes categorías

Primero haremos de nuevo la gráfica de densidad. Para diferenciar las categorías necesitaremos varios colores, pero con transparencia, y para una mejor visualización usaremos la escala logarítmica.

Crearemos primero nuestro lienzo y luego la geometría.

```
gg_bi <-base_descriptiva %>%
  filter(clase2==1 & pos_ocu!=4) %>% # Los que no reciben ingresos
  mutate(pos_ocu=as_label(pos_ocu)) %>% ## Etiqueta variable categórica
  ggplot(aes(x=log(ingocup2))) # Dibuja el lienzo en el eje de las x
gg_bi
```

**Gráfica iv-6. Lienzo para los ingresos por hora modificado**

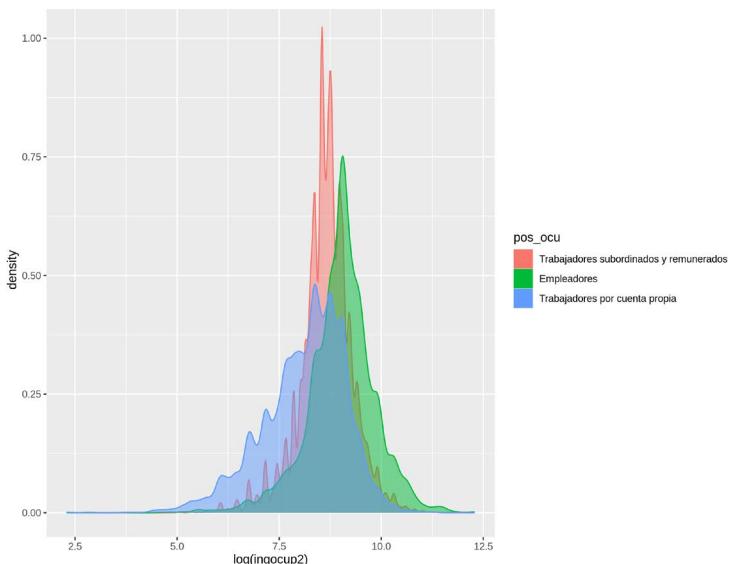


FUENTE: elaboración propia con datos de la ENOE.

Una vez que tenemos nuestro lienzo, de forma aditiva agregamos nuestras geometrías. En este caso introduciremos la variable tanto para colorear la forma como el relleno (“fill”). La opción “alpha” sirve para modificar las transparencias de nuestro relleno. Si el valor se acerca a 1, tendremos un relleno más opaco; si ponemos un valor más cercano a 0, será más transparente.

```
gg_bi + geom_density(aes(
  fill=pos_ocu, # el relleno tendrá un color por categoría
  color=pos_ocu, # también el color de la línea
  alpha=I(.5)
) # pondremos un color transparente
)
```

**Gráfica iv-7. Distribución del logaritmo de los ingresos mensuales, según posición en la ocupación. México, trimestre III, 2019**



FUENTE: elaboración propia con datos de la ENOE.

Es evidente que las distribuciones de los logaritmos de los ingresos de estos tres grupos son diferentes. Vemos cómo la de los empleadores está mucho más a la derecha, por sus mayores niveles de ingresos; mientras que los trabajadores por cuenta propia están más a la izquierda, pero también tienen una forma más achataada que demuestra que se trata de un grupo con más heterogeneidad. Los subordinados tienen una distribución mucho más “delgada”, lo que significa una varianza mucho menor en los ingresos.

Esta gráfica también se puede replicar con el paquete *esquisse*. Incluso una gráfica de caja y brazos podría ser muy informativa y aportaría mayor visibilidad a los valores atípicos.

## b. Estadísticos para grupos

Al igual que en el análisis univariado es útil graficar, pero también calcular las medidas numéricas. Con el fin de hacer estadísticos para

grupos, podemos agregar a los códigos anteriores una línea intermedia en los *pipes* que hemos utilizado anteriormente. En esta línea le ordenamos que nos haga las operaciones de manera agrupada con “group\_by”. Esto debe hacerse antes de que declaremos el “sumarise” que queremos realizar. Esto lo desarrollamos en los siguientes códigos:

```
base_descriptiva %>%
  filter(clase2==1 & pos_ocu!=4) %>% #vamos a quitar Los que no reciben ingresos
  group_by(as_label(pos_ocu)) %>% # hace el agrupamiento para la variables categóricas
  summarise(media=mean(ingocup2, na.rm=T), # checa que ponemos que nos remueva los missings
            sd=sd(ingocup2, na.rm=T),
            mediana=median(ingocup, na.rm=T))

## # A tibble: 3 x 4
##   `as_label(pos_ocu)`      media     sd mediana
##   <fct>                <dbl>    <dbl>  <dbl>
## 1 Trabajadores subordinados y remunerados 7028.  5556.   5000
## 2 Empleadores             11487. 12255.   5000
## 3 Trabajadores por cuenta propia       5196.  5860.   2150
```

Este resumen nos brinda las medidas centrales de cada grupo y la desviación estándar. Es importante observar que dentro de las funciones de cálculo para los estadísticos utilizados, hemos incluido un nuevo argumento “na.rm=T”; ello nos dice que en el cálculo no se toman en cuenta los valores perdidos. De otro modo, por coerción, tendríamos como resultado valores “NA”.

### c. Estadísticos con datos expandidos

Nuevamente, las estimaciones de la muestra pueden diferir de acuerdo con el factor de expansión, que es parte del diseño muestral. *R* cuenta con una función base que calcula la media expandida o ponderada:

```
base_descriptiva %>%
  filter(clase2==1 & pos_ocu!=4) %>% #vamos a quitar Los que no reciben ingresos
  group_by(as_label(pos_ocu)) %>% # hace el agrupamiento para la variables categóricas
  summarise(media=mean(ingocup2, na.rm=T),
            media_ponderada=
              weighted.mean(ingocup2, na.rm=T, w=fac)) # checa que ponemos que nos
remueva Los missings

## # A tibble: 3 x 3
##   `as_label(pos_ocu)`      media media_ponderada
##   <fct>                <dbl>        <dbl>
## 1 Trabajadores subordinados y remunerados 7028.      6577.
## 2 Empleadores             11487.     10602.
## 3 Trabajadores por cuenta propia       5196.      4846.
```

Vemos que la media ponderada es menor. Por lo que es importante tener en cuenta la estructura de la población expandida cuando hacemos referencia a la población.

El paquete *svyvr* es útil para otras medidas y también tiene capacidad de introducir el diseño muestral completo:

```
base_descriptiva %>%
  as_survey_design(weights = fac) %>% #establece los pesos
  mutate(pos_ocu=as_label(pos_ocu)) %>% #para mejor lectura de las etiquetas
  filter(clase2==1 & pos_ocu!=4) %>% #vamos a quitar los que no reciben ingresos
  group_by(pos_ocu) %>% # hace el agrupamiento para las variables categóricas
  summarise(media_ponderada=
    survey_mean(ingocup2, na.rm=T))

## # A tibble: 4 x 3
##   pos_ocu                         media_ponderada media_ponderada...
##   <fct>                               <dbl>                <dbl>
## 1 Trabajadores subordinados y remunerados      6577.               26.1
## 2 Empleadores                           10602.              198.
## 3 Trabajadores por cuenta propia        4846.               47.8
## 4 Trabajadores sin pago                      0                  0
```

Este paquete también calcula el error estándar de nuestra estimación. En el siguiente capítulo veremos qué tan importante es este valor para nuestra inferencia de la población.

Para estimar la mediana ponderada:

```
base_descriptiva %>%
  as_survey_design(weights = fac) %>% #establece los pesos
  mutate(pos_ocu=as_label(pos_ocu)) %>% #para mejor lectura de las etiquetas
  filter(clase2==1 & pos_ocu!=4) %>% #vamos a quitar los que no reciben ingresos
  group_by(pos_ocu) %>% # hace el agrupamiento para las variables categóricas
  summarise(mediana_ponderada=
    survey_median(ingocup2, na.rm=T)) # mediana ponderada

## # A tibble: 4 x 3
##   pos_ocu                         mediana_ponderada mediana_ponderada...
##   <fct>                               <dbl>                <dbl>
## 1 Trabajadores subordinados y remune...      5547                 54.8
## 2 Empleadores                           8333                 153.
## 3 Trabajadores por cuenta propia        3440                 110.
## 4 Trabajadores sin pago                      0                  0
```

Para estimar los valores asociados a percentiles y otras medidas de posición:

```
base_descriptiva %>%
  as_survey_design(weights = fac) %>% #establece los pesos
  mutate(pos_ocu=as_label(pos_ocu)) %>% #para mejor lectura de las etiquetas
  filter(clase2==1 & pos_ocu!=4) %>% #vamos a quitar los que no reciben ingresos
  group_by(pos_ocu) %>% # hace el agrupamiento para las variables categóricas
  summarise(mediana_ponderada=
    survey_quantile(ingocup2, c(0.25, 0.75), na.rm=T)) # Nos da q1 y q3
```

```
## # A tibble: 4 x 5
##   pos_ocu mediana_pondera... mediana_pondera... mediana_pondera...
##   <dbl>      <dbl>          <dbl>          <dbl>
## 1 Trabaj...     3870        7740           0
## 2 Emplea...     5000       12900         168.
## 3 Trabaj...     1600        6450        54.8
## 4 Trabaj...        0           0           0
## # ... with 1 more variable: mediana_ponderada_q75_se <dbl>
```

Es importante recordar que todos los resultados se pueden guardar en un objeto para volver a ellos o graficarlos.

Es relevante mencionar que, al expandir los datos con el factor de expansión, hay otros elementos del diseño muestral que pueden incorporarse y robustecer nuestros resultados. Proponemos consultar el material desarrollado por Martínez Sánchez (2017).

## D. El coeficiente de Gini

En esta sección queremos introducir una medida de desigualdad muy utilizada: el coeficiente de Gini. Este coeficiente es denominado así por su creador, Corrado Gini, en 1921. Normalmente se utiliza para medir desigualdades económicas, como las distribuciones de ingresos y de riqueza.

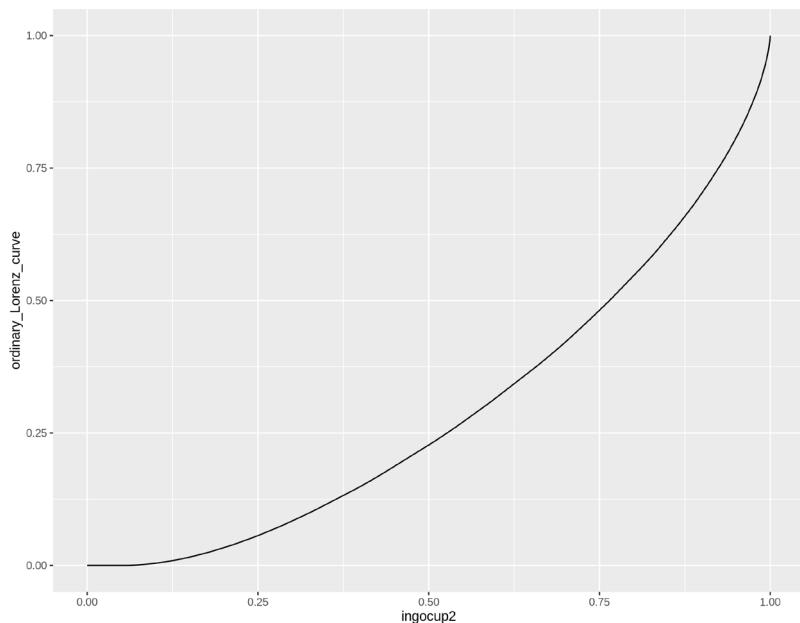
El coeficiente varía de 0 (o 0%) a 1 (o 100%), donde 0 representa la igualdad perfecta y 1 representa la desigualdad perfecta.

El coeficiente proviene de otra herramienta, la curva de Lorenz, creada por Max O. Lorenz en 1905. Esta curva representa la proporción del ingreso total de la población (eje y) que el x% inferior de la población gana acumulativamente. La línea a 45 grados representa la igualdad perfecta de ingresos. Para profundizar sobre este tema, léase Dorfman (1979) y Milanovic (1997).

Primero vamos a hacer la curva de Lorenz con ayuda del paquete *gglorenz*, que al estar basado en *ggplot2* es completamente compatible con nuestros lienzos:

```
base_descriptiva %>%
  filter(clase2==1) %>%
  ggplot(aes(ingocup2)) + stat_lorenz() # esta geometría sólo se puede usar con el
paquete gglorenz
```

**Gráfica iv-8. Curva de Lorenz de los ingresos mensuales válidos.  
México, trimestre III, 2019**

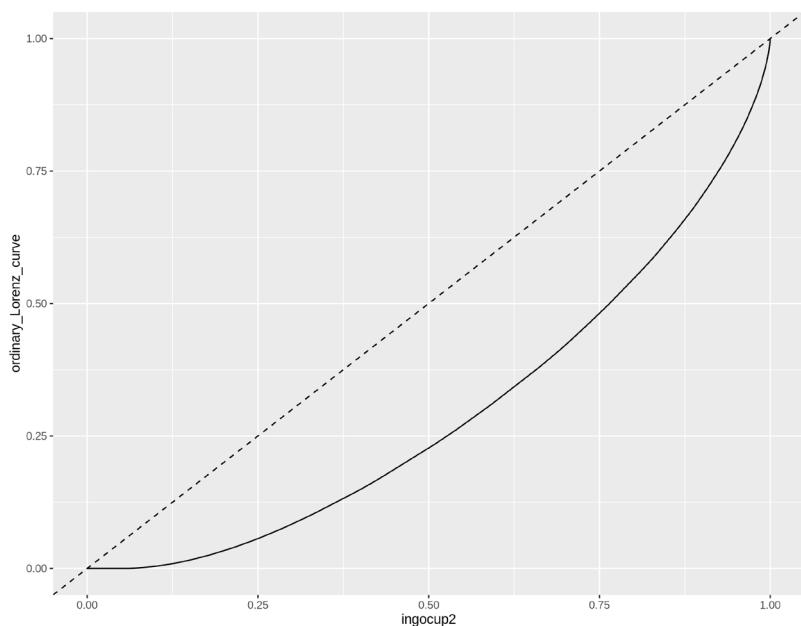


FUENTE: elaboración propia con datos de la ENOE.

En el eje de las  $x$  se aprecia cuándo se acumula cierto porcentaje de la población al sumar por orden su nivel de ingresos. Así, en 0.25, tenemos al 25% inferior de la población. Vemos que acumulan mucho menos del 25% de los ingresos, de ahí que nos muestre la desigualdad de los ingresos. Podemos agregar fácilmente la línea de comparación de 45 grados que nos compara con una situación hipotética de igualdad:

```
base_descriptiva %>%
  filter(clase2==1) %>%
  ggplot(aes(ingocup2)) +
  stat_lorenz() + # esta geometría sólo se puede usar con el paquete gglorenz
  geom_abline(linetype = "dashed") # agrega una línea de 45 grados punteada
```

**Gráfica iv-9. Curva de Lorenz y línea de igualdad de los ingresos mensuales válidos. México, trimestre III, 2019**



FUENTE: elaboración propia con datos de la ENOE.

El coeficiente de Gini es igual al área debajo de la línea de igualdad perfecta (0.5 por definición) menos el área debajo de la curva de Lorenz, dividida por el área debajo de la línea de igualdad perfecta. En otras palabras, es el doble del área entre la curva de Lorenz y la línea de igualdad perfecta. El valor del coeficiente de Gini se puede calcular con el paquete *ineq* y la función del mismo nombre. Es un paquete muy potente para analizar la desigualdad con diferentes medidas, así como la pobreza.

```
base_descriptiva %>%
  filter(clase2==1) %>%
  with(ineq(ingocup2, #variable de ingresos
            type="Gini", #tipo de medida
            na.rm=T)) # para que remueva los datos faltantes

## [1] 0.4125001
```

## E. La relación entre dos variables cuantitativas

Por el momento, hemos realizado un análisis descriptivo para una sola variable cuantitativa a lo largo de las categorías de una variable cualitativa. ¿Qué sucede cuando tenemos dos variables cuantitativas? Como con el análisis univariado, tendremos gráficas y medidas numéricas.

### a. Gráfica de dispersión

Para visualizar la relación entre dos variables podemos hacer una gráfica donde el eje de las  $y$  sean los ingresos en escala logarítmica y el eje de las  $x$  sean los años de escolaridad. Es decir, un plano cartesiano donde las coordenadas refieren a las variables en análisis y se dibuja un punto por observación.

Primero, revisemos la variable de escolaridad, ya que la de ingresos la hemos analizado bien:

```
base_descriptiva %>%
  with(summary(anios_esc))

##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.
##      0.000   3.000   9.000   7.953  12.000   99.000
```

Observamos que hay valor máximo de 99, lo cual es imposible porque nadie puede estudiar 99 años en una vida, al menos, humana. Si revisamos el material metodológico de la ENOE, sabemos que se trata de valores no especificados, por lo que habrá que recodificar como lo hicimos con los ingresos:

```
base_descriptiva<-base_descriptiva %>%
  mutate(anios_esc2=
    ifelse( # función ifelse
      anios_esc==99, #si cumple con esta condición
      NA, #toma este valor
      anios_esc)) # si no, toma éste

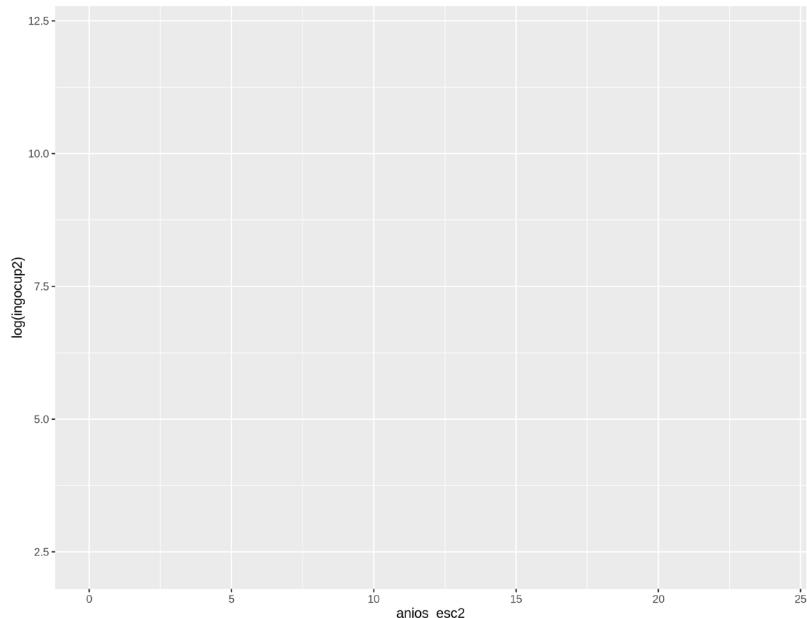
base_descriptiva %>%
  with(summary(anios_esc2))

##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.    NA's
##      0.000   3.000   9.000   7.845  12.000  24.000      473
```

Una vez que tenemos nuestras variables hacemos la gráfica como siempre, con “`ggplot()`”, iniciando con el lienzo. Recuerda que esto también lo puedes hacer con el complemento de *esquisse*.

```
scatter0<-base_descriptiva %>%
  filter(clase2==1) %>%
  ggplot(aes(y=log(ingocup2), x=anios_esc2))
scatter0
```

**Gráfica iv-10. Lienzo bivariado. En el eje de las x están los años de escolaridad y en el de las y los logaritmos de los ingresos válidos**

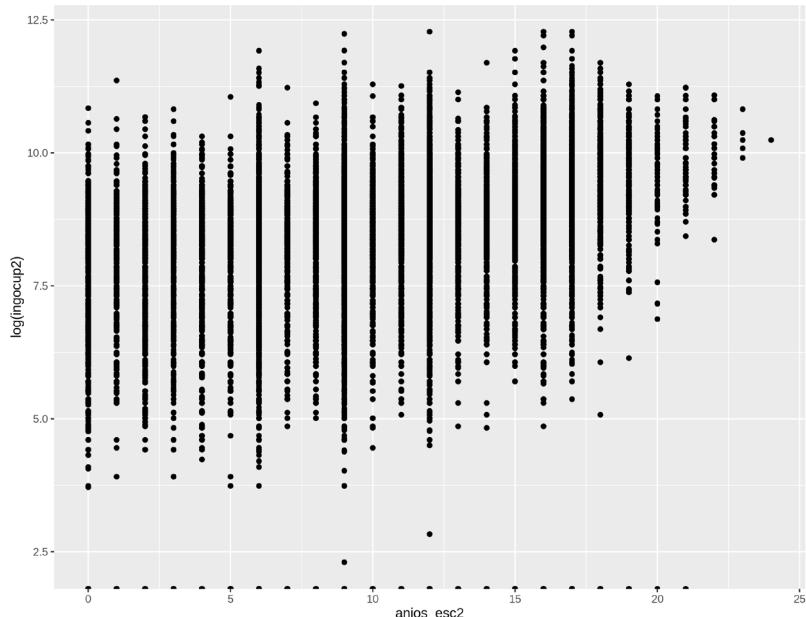


FUENTE: elaboración propia con datos de la ENOE.

Una vez hecho esto, añadimos la geometría:

```
scatter0 + geom_point() # dibuja un punto en la coordenada  
## Warning: Removed 44948 rows containing missing values (geom_point).
```

**Gráfica iv-11. Diagrama de dispersión de la relación entre los ingresos y los años de escolaridad. México, trimestre III, 2019**

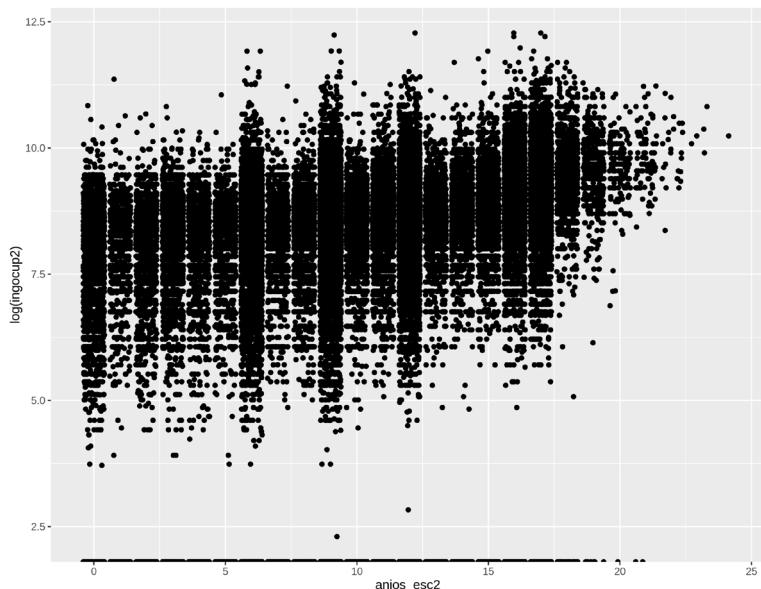


FUENTE: elaboración propia con datos de la ENOE.

La geometría *jitter* desagrupa los puntos superpuestos lo que puede mejorar muchas veces nuestra gráfica, sobre todo cuando tenemos una gran cantidad de observaciones en análisis:

```
scatter0 + geom_jitter() # dibuja un punto en la coordenada
```

**Gráfica iv-12. Diagrama de dispersión de la relación entre los ingresos y los años de escolaridad. México, trimestre III, 2019.  
Con la opción “jitter()”**



FUENTE: elaboración propia con datos de la ENOE.

Con esta gráfica podemos ver una relación. De acuerdo con Moore (2010, p. 105), para describir el aspecto general de un diagrama de dispersión se necesita analizar: la forma, la dirección y la fuerza de la relación; así como las observaciones atípicas, es decir, valores individuales que quedan fuera del aspecto general de la relación. Además de ello, se puede analizar si hay grupos de datos —que a veces se denominan clústers— que tienen comportamientos particulares.

En este caso particular, la forma es ligeramente curvilínea pues no parece una línea recta; más bien aparece como estancada en los primeros años de escolaridad y se vuelve más inclinada a partir del año 12. Es ascendente pues parece que a mayor educación (mayores los niveles de escolaridad) hay más ingresos mensuales, en escala logarítmica. La fuerza de la relación se puede decir que no es tan fuerte, es decir, que no se ve tan definida a través de los puntos.

¿Pero qué es fuerte? Sin duda, quisiéramos tener alguna medida numérica de esta fuerza. Para el caso de una línea recta, podemos establecer una medida de correlación lineal.

## b. Correlación

La correlación mide la fuerza y la dirección de la relación lineal entre dos variables cuantitativas. Se abrevia como  $r$  para la muestra y como  $\rho$  para la población.

Se obtiene de la siguiente fórmula:

$$r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

Es decir, es la sumatoria del producto de los valores estandarizados entre las dos variables. El coeficiente de Pearson,  $r$ , puede tener valores entre  $-1$  y  $1$ .

Para referencia:

- $1$  es una correlación positiva perfecta, una línea recta de  $45$  grados.
- $0$  es sin correlación (los valores no parecen vinculados en absoluto).
- $-1$  es una correlación negativa perfecta.

El coeficiente de correlación,  $r$ , no hace ninguna distinción entre las variables explicativas y dependientes. Y en el cálculo de la correlación, tampoco hace ninguna diferencia entre cuál variable se llama  $x$  y cuál  $y$ .

El coeficiente  $r$  utiliza los valores estandarizados de las observaciones. Además, no cambia cuando cambiamos las unidades de medida de  $x$ ,  $y$ , o ambos. Si pasáramos los pesos a dólares y los años de escolaridad a minutos, el cálculo de esa correlación no cambiaría. La correlación  $r$ , en sí, no tiene unidad de medida.

Para calcular  $r$  con nuestras variables, tenemos el siguiente código, usando el comando “`cor()`”:

```
base_descriptiva %>%
  filter(clase2==1 & ingocup2>0) %>% # EL Logaritmo de 0 no es número, por eso Los
  eliminamos
  with(cor(x=anios_esc2, # variable 1
           y=log(ingocup2), # variable 2
           use="pairwise" ))
## [1] 0.3833647
```

La correlación es de 0.38, lo cual, si bien no es tan cercano a 0, no se acerca a 1; por lo tanto, decimos que es una correlación *lineal* no tan fuerte. Es importante notar que  $r$  sólo mide la fuerza de una relación lineal entre dos variables, no describe las relaciones curvilíneas entre variables, aunque sean muy fuertes.

En este capítulo hemos avanzado en el análisis del comportamiento de los ingresos de los trabajadores en México. Sin duda, un sesgo importante hacia la derecha y gran presencia de valores atípicos dan cuenta de una gran desigualdad, donde hay muy pocos trabajadores con ingresos muy altos y una gran mayoría con ingresos muy bajos. También se evidenció, a partir de la comparación de las distribuciones y las medidas numéricas, las diferencias en los desempeños entre la posición por ocupación. Es notable que quienes trabajan por cuenta propia tienen mayor dispersión, es decir, son más heterogéneos en sus ingresos y, además, ganan menos en promedio.

También dimos cuenta de la desigualdad a partir del cálculo del coeficiente de Gini, con un valor de 0.41. Se trata de un valor alejado de 0, que sería una situación de igualdad perfecta. Todo ello abona a que sigamos encontrando elementos para denotar que las desigualdades se encuentran muy presentes en el mercado de trabajo mexicano.

Finalmente, exploramos la relación de los ingresos en términos de los niveles de escolaridad: encontramos una relación positiva, aunque no tan fuerte. Vamos a seguir explorando esta relación en el capítulo vi, donde además discutiremos no sólo sobre la fuerza de la relación lineal, sino también podremos complejizar el análisis, estableciendo su significancia estadística e incorporando más variables.

# V. Introducción a la inferencia

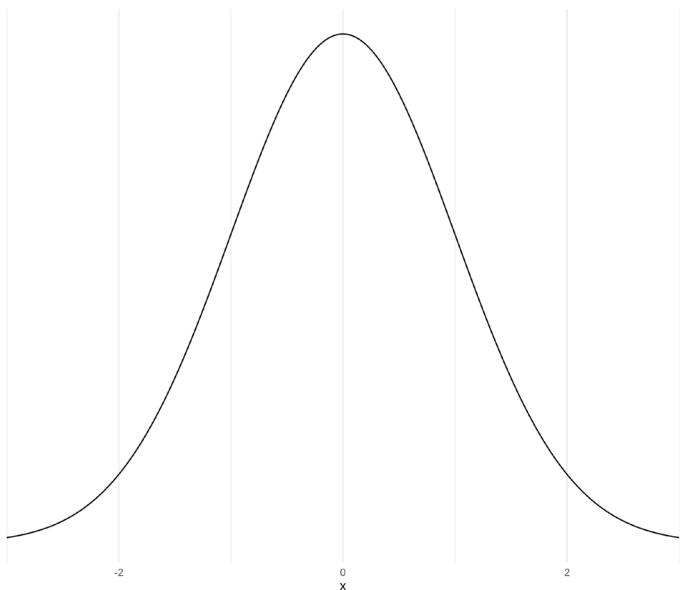
## Introducción

El proceso de inferencia consiste en estimar parámetros poblacionales con información de la muestra, pero para ello se necesitan algunos requisitos. Precisamos que nuestra información provenga de un muestreo aleatorio, con el fin de eliminar sesgos de información y poder aplicar las leyes matemáticas asociadas a la probabilidad, tales como el teorema del Límite Central y la Ley de los Grandes Números. Ello nos permite establecer que la población se distribuye como una normal.

Para comprender mejor este capítulo, quien lo lee debe retomar conceptos de probabilidad y el estudio de algunas distribuciones. De nuevo, se recomienda revisar estos elementos en los textos de Moore (2010) y de Mendenhall *et al.* (2014).

Iniciemos recordando sobre la función de densidad de probabilidad, la distribución normal. En el eje de las  $x$  irán los valores de nuestra variable aleatoria (pues debe provenir de un proceso aleatorio, como el experimento de seleccionar una muestra). En el eje de las  $y$ , tenemos el valor de la probabilidad asociado a diferentes niveles posibles de  $x$ . Esta distribución normal se puede graficar utilizando *R*. Esto es un poco más avanzado de lo que veníamos trabajando; los códigos son los siguientes:

```
ggnormal <- ggplot(data = data.frame(x = c(-3, 3)), aes(x)) +  
  stat_function(fun = dnorm, n = 1000, args = list(mean = 0, sd = 1)) + ylab("") +  
  scale_y_continuous(breaks = NULL)  
ggnormal + theme_minimal()
```

**Gráfica v-1. Distribución normal estándar**

Una distribución normal tiene dos parámetros fundamentales: la media  $\mu$  y la desviación estándar  $\sigma$ . Decimos que una variable se distribuye como una normal de esta forma  $x \sim N(\mu, \sigma)$ . Como se observa, es una distribución simétrica. En este caso, graficamos una distribución normal centrada en 0, es decir, su media es 0 y su desviación estándar es 1. Dicho de otra forma, nuestra variable  $x \sim N(0,1)$  —esta distribución— se llama estándar y se puede definir con una  $Z$  de la siguiente manera:  $x \sim Z$ .

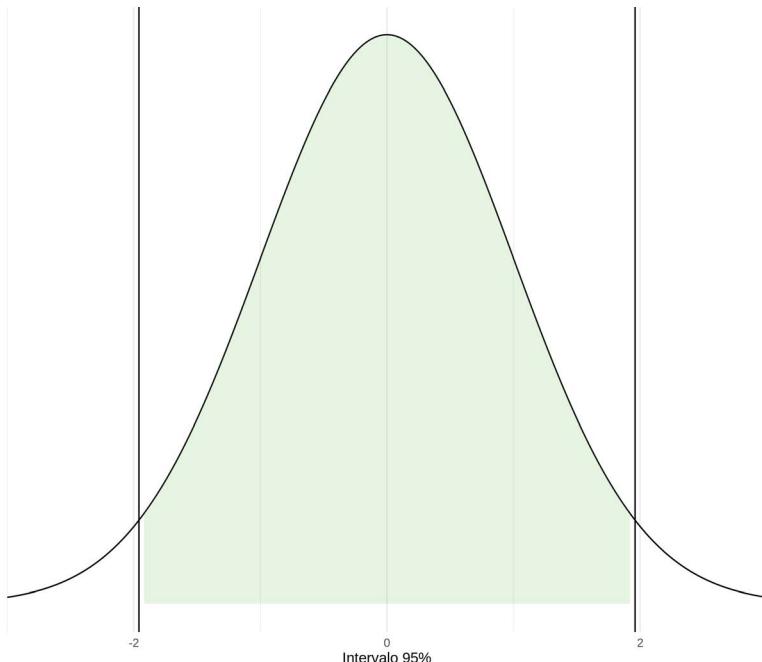
Los valores  $Z$  de una distribución estándar nos dan valores de probabilidad. Al ser una distribución de probabilidad continua, no podremos encontrar valores específicos, pero podemos calcular áreas debajo de la curva. Esto es parecido a integrar una función. Por ejemplo, el área debajo de la curva entre los valores  $Z$  de  $-1.96$  y  $1.96$ , representan una probabilidad de 0.95.

Esto lo podemos comprobar con el comando “`pnorm()`”, que nos da la probabilidad acumulada (desde  $-\infty$  hasta el valor solicitado en el argumento). Por tanto, para hallar el área entre dos valores, habrá una resta de esta función evaluada entre ellos:

```
pnorm(1.96) - pnorm(-1.96)
```

```
## [1] 0.9500042
```

**Gráfica v-2. Distribución normal estándar, con un área debajo de la curva de 0.95 central**



A lo largo de los capítulos hemos hablado un poco de la importancia de contar con estimaciones para la población. En el capítulo anterior, creamos un par de variables para nuestro análisis, por lo que puedes descargar el ambiente desde este enlace: <<https://tinyurl.com/Base-Enoe>>.

```
if (!require("pacman")) install.packages("pacman") # instala pacman si se requiere
## Loading required package: pacman

#cargan los paquetes necesarios para la práctica de este capítulo
pacman::p_load(tidyverse, haven, janitor, broom, svyr, RColorBrewer, wesanderson,
sjlabelled, DescTools)
load("base_descriptiva2.RData") ## Recuerda tenerla descargada en tu carpeta de trabajo
```

Nuestro primer acercamiento a la inferencia vendrá desde análisis univariados y bivariados. En el próximo capítulo ahondaremos más en los modelos y, por tanto, podremos tener análisis más interesantes con más variables.

Recomiendo revisar los conceptos de probabilidad y de distribuciones de probabilidad con profundidad en libros estadísticos, puesto que son conceptos que requieren mucho cuidado y tienen elementos matemáticos fundamentales. Este texto no podría aportar más de lo que se encuentra en los libros que ya se han recomendado en la introducción. En este capítulo pondremos ejemplos de cómo aplicarlos directamente a la base de datos de la ENOE.

En general, para las estimaciones poblacionales, tendremos un estadístico muestral que se aproxima al parámetro poblacional, más o menos un error. Ello da como resultado un intervalo de confianza a un nivel por determinar.

$$\text{parámetro} = \text{estadístico} \pm \text{error}$$

Este error es función de la confiabilidad con la que queremos hacer la estimación, la cual se asocia con la probabilidad. Mientras que también el error toma en algunas ocasiones la heterogeneidad de la población y el tamaño de la muestra, esta estimación puede verse como un intervalo. Normalmente construimos intervalos al 95% de confianza.

La segunda forma de inferencia más usada son las pruebas de significación o de hipótesis. El objetivo de ellas es valorar la evidencia proporcionada por los datos a favor de alguna hipótesis sobre la población. En el caso de las pruebas de hipótesis, dado un valor específico asumido para el parámetro a estimar, revisamos cuál sería la probabilidad de que nosotros obtengamos valores más extremos al estadístico (muestral).

A continuación establecemos cómo estimar algunos parámetros poblacionales más comunes, los cuales ya revisamos en capítulos anteriores para el caso de la estadística descriptiva. Por tanto, revisamos la inferencia para la media de una muestra y la diferencia de dos muestras. Posteriormente, revisamos las estimaciones de

la varianza y la diferencia de las varianzas. Además, se incluye la prueba estadística chi-cuadrado ( $\chi^2$ ) de independencia, así como la prueba  $F$  de análisis de varianza para establecer el efecto de un factor en una variable.

## A. La media poblacional $\mu$

Para la estimación de una media poblacional, tendremos las siguientes fórmulas, dependiendo del tamaño de la muestra.

Para muestras grandes:

$$\mu = \bar{x} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Donde  $\bar{x}$  es estimador muestral de la media poblacional  $\mu$ ,  $Z$  es el valor asociado a la confiabilidad, asumiendo una distribución normal estandarizada.

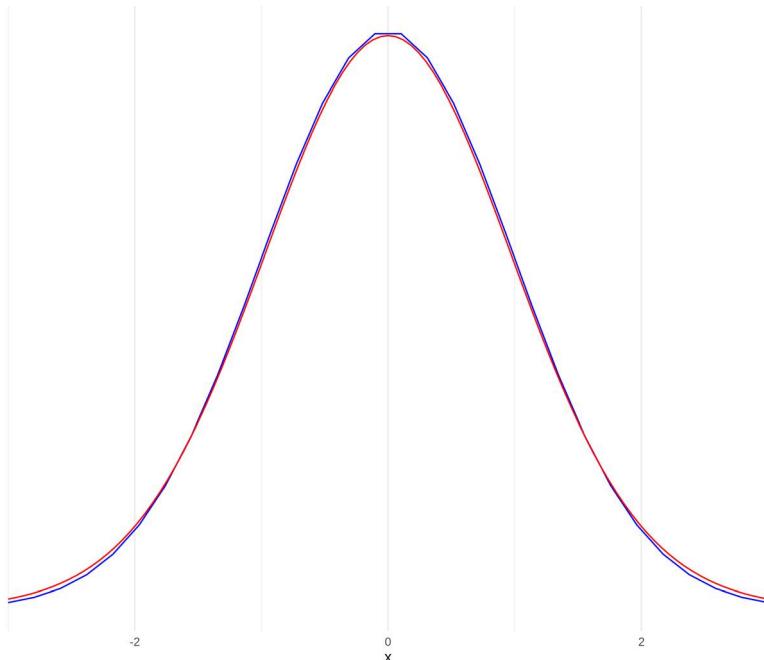
Es importante señalar que el valor de  $Z$  más común es 1.96, que se asocia al 95% de confiabilidad en los intervalos, correspondiente al valor de  $(1 - \alpha) * 100$  cuando  $\alpha = 0.05$ , de ahí que lo hayamos presentado en la Gráfica v-2.  $s$  es la desviación estándar y  $n$  es el tamaño de la muestra. Si conociéramos la desviación estándar de la población, usaríamos  $\sigma$  en lugar de  $s$ .

En el caso de muestras pequeñas, utilizamos la distribución  $t$ , la cual necesita establecer el número de grados de libertad:

$$\mu = \bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

con  $n - 1$ , grados de libertad.

La distribución  $t$ , de Student, es una distribución que se parece mucho a la normal, sólo que con colas más amplias. Cuando tenemos muestras grandes, ésta coincide con la normal. De ahí que los paquetes estadísticos tengan un solo comando para estimar la media poblacional. Aquí se ve cómo se superponen estas distribuciones con una muestra igual a 30 ( $n = 30$ ). En azul, podemos observar la distribución normal y, en rojo, la  $t$ -Student.

**Gráfica v-3. Distribución normal estándar y distribución t-Student**

En el caso de *R* se utiliza la función “*t.test()*”. Este comando sirve para calcular diferentes tipos de pruebas estadísticas que tienen como base la distribución t que, como mencionamos, se aproxima a una normal cuando las muestras son grandes.

Si queremos hacer estimaciones para nuestra variable “*ingocup*”, usamos el siguiente código:

```
base_descriptiva %>%
  with(t.test(ingocup2))

##
##  One Sample t-test
##
##  data:  ingocup2
##  t = 381.08, df = 134510, p-value < 2.2e-16
##  alternative hypothesis: true mean is not equal to 0
##  95 percent confidence interval:
##    6419.223 6485.597
##  sample estimates:
##  mean of x
##    6452.41
```

Al observar los resultados tenemos que, para el tercer trimestre de 2019, los trabajadores en México ganaron, con 95% de confianza, entre 6 419.22 y 6 485.60 pesos. Para manejar y consultar estos resultados, los guardamos en un objeto:

```
t.test0<-base_descriptiva %>%
  with(t.test(ingocup2))
tidy(t.test0)

## # A tibble: 1 x 8
##   estimate statistic p.value parameter conf.low conf.high method
##     <dbl>     <dbl>    <dbl>      <dbl>     <dbl>    <dbl> <chr>
## 1 6452.     381.     0     134510     6419.    6486. One S...
## # ... with 1 more variable: alternative <chr>
```

La función “tidy()” hace que el resultado se vuelva un tibble, es decir, una tabla muy compatible con el *tidyverse*. Esto puede ser útil cuando queremos comparar estimaciones.

De forma predeterminada, el programa asume una prueba de hipótesis donde el valor objetivo es cero. Lo cual se escribe de la siguiente manera:

$$H_o: \mu = 0$$

$$H_a: \mu \neq 0$$

Esto lo podemos modificar. Probablemente sea mucho más interesante comparar contra valores normativos. Por ejemplo, la línea de bienestar de CONEVAL (2020) calcula que la línea de pobreza por ingreso (valor de la canasta alimentaria y no alimentaria) fue de 3 085 pesos por persona. Comparemos entonces contra un valor que representa ingresos laborales que le permita al trabajador o trabajadora y otra persona de su hogar cumplir con esta canasta. Es decir, planteamos las hipótesis:

$$H_o: \mu = 2 * 3085$$

$$H_a: \mu \neq 2 * 3085$$

```
base_descriptiva %>%
  with(t.test(ingocup2, mu=2*3085))

##
##  One Sample t-test
##
```

```
## data: ingocup2
## t = 16.679, df = 134510, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 6170
## 95 percent confidence interval:
## 6419.223 6485.597
## sample estimates:
## mean of x
## 6452.41
```

Según los resultados obtenidos, observamos que la media poblacional, de ser igual a 6 170, como establece la hipótesis nula, estaría a 16.93 errores estándar de la media calculada (6 452.41), lo que nos da el estadístico de prueba  $t$ . Es decir, es un valor muy lejano a la media calculada en términos de la distribución, por lo cual habría poca probabilidad de que represente la media poblacional y ello se muestra en el valor- $p$  tan pequeño. En ese sentido, rechazamos la hipótesis nula de que los ingresos laborales son iguales a dos veces la línea de pobreza, por lo que podemos afirmar que los ingresos laborales de la población son distintos a ese valor.

Por supuesto que sería más útil saber si los ingresos son mayores o menores. Las anteriores han sido pruebas de dos colas o bilaterales. Es decir, evalúan si un valor es distinto a otro, sin importar si son mayores o menores. Esto lo podemos cambiar añadiendo el código “alternative”, que establece la dirección de la prueba estadística. Por ejemplo, si se quisiera evaluar la hipótesis nula de que la  $\mu$  es mayor o igual al valor objetivo contra la alternativa de si en la población los ingresos son menores a este valor:

$$H_o: \mu \geq 2 * 3085$$

$$H_a: \mu < 2 * 3085$$

En el código establecemos:

```
base_descriptiva %>%
  with(t.test(ingocup2, mu=2*3085, alternative="less"))

##
## One Sample t-test
##
## data: ingocup2
## t = 16.679, df = 134510, p-value = 1
## alternative hypothesis: true mean is less than 6170
## 95 percent confidence interval:
## -Inf 6480.261
## sample estimates:
## mean of x
## 6452.41
```

Como se ve, la probabilidad casi es unitaria y, si bien el valor del estadístico de prueba no cambia, no rechazamos la hipótesis de que los ingresos son mayores o iguales a 6 170. Por tanto, no podemos afirmar que los trabajadores en México tienen ingresos inferiores a dos veces la línea de ingresos de pobreza.

Generalmente, el valor-*p* de una prueba de hipótesis se compara contra un nivel de significancia. El valor estándar y adoptado por la comunidad es 0.05; si el valor-*p* es menor a este valor, rechazamos la hipótesis nula y afirmamos la alternativa.

Este comando también calcula un intervalo de confianza de una sola cola, es decir, en lugar de tener dos valores entre los cuales tenemos a nuestro parámetro, tenemos el valor máximo que puede obtener nuestra estimación. En este caso, con 95% de confianza, podemos decir que lo más que gana un trabajador mexicano son 6 480.26 pesos.

Las pruebas de hipótesis y el nivel de significancia han tenido su discusión en la comunidad estadística y científica, por lo que hay que tener cuidado cuando hablamos de resultados significativos de una prueba de hipótesis, sobre todo si no revisamos los intervalos de confianza. Se sugiere sobre ello la lectura meticulosa de Amrhein, Greenland & McShane (2019), así como lo que estableció la Asociación Estadounidense de Estadística (ASA, por sus siglas en inglés) (Wasserstein & Lazar, 2016).

## B. Diferencia de medias de dos grupos

Muchas veces queremos comparar cómo se comporta la media entre dos grupos y si lo que observamos como diferencia entre las muestras de cada grupo da cuenta de lo que sucede entre los grupos de la población. Para ello, veamos qué nos dicen los datos acerca de los ingresos diferenciados entre hombres y mujeres.

Para estimar estos elementos, en el caso de muestras grandes, se sigue la fórmula:

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Donde  $(\bar{x}_1 - \bar{x}_2)$  es el estimador de la muestra de la diferencia de las medias;  $Z$ , de nuevo, es valor asociado a la confiabilidad, asumiendo una distribución normal estandarizada;  $s_1$  y  $s_2$  son las desviaciones estándar; así como  $n_1$  y  $n_2$  son los tamaños de las muestras de cada grupo. Al igual que para el caso de una sola muestra, en caso de que conociéramos las desviaciones estándar de las poblaciones, usaríamos  $\sigma_1$  y  $\sigma_2$ , en lugar de  $s_1$  y  $s_2$ .

Para muestras chicas, lo anterior también implica suposiciones sobre las varianzas de las poblaciones. Esto sólo se menciona en la sección del proceso. Dado que la ENOE no es una muestra chica, no es objetivo directo de este texto; no obstante, se puede consultar el proceso al final de esta sección.

Primero, se puede estimar cómo se diferencian los ingresos a nivel muestral, utilizando el mismo comando que hemos utilizado en capítulos anteriores, “`summarise()`”, tal como sigue:

```
base_descriptiva %>%
  group_by(as_label(sex)) %>%
  summarise(avg_ing = mean(ingocup2, na.rm=T))

## # A tibble: 2 x 2
##   `as_label(sex)` avg_ing
##   <fct>           <dbl>
## 1 Hombre            7247.
## 2 Mujer             5337.
```

Como vemos, los hombres ganan más que las mujeres con una diferencia de casi 2 000 pesos mensuales. ¿Esto se mantendrá para la población? Para ello, volvemos a utilizar el comando “`t.test()`” que ya sabemos que, al operar con muestras grandes, sus valores de este test se aproximan a la distribución normalizada  $Z$ . Es importante anotar que debemos colocar la variable cuantitativa seguida del símbolo `~` y colocar la variable que divide a nuestra población en dos grupos. El programa dará error si identifica más de dos muestras.

```
base_descriptiva %>%
  with(t.test(ingocup2~sex))

##
## Welch Two Sample t-test
##
## data: ingocup2 by sex
## t = 58.58, df = 133434, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1845.228 1972.977
## sample estimates:
## mean in group 1 mean in group 2
##           7246.560          5337.457
```

El grupo 1 queda definido por los hombres y el grupo 2 por las mujeres, pues éste es el orden de las categorías en la variable; en relación con ese orden hay que ser cuidadosos e identificar bien lo que define cada grupo. Lo mostrado en este caso denota que la diferencia de ingresos laborales estaría, con 95% de confianza, entre 1 845.23 y 1 972.98 pesos.

En términos de las hipótesis, éstas se plantean de la siguiente forma:

$$H_o: \mu_1 - \mu_2 = 0$$

$$H_a: \mu_1 - \mu_2 \neq 0$$

Como vemos aquí, la prueba de hipótesis, que de forma determinada tiene el programa, cobra más sentido pues una diferencia, al compararla con cero, la estamos comparando contra la igualdad de medias. Podemos reordenar nuestras hipótesis de esta manera:

$$H_o: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

Por lo que, al tener un valor-*p* muy pequeño, podemos rechazar la hipótesis de igualdad de medias y por tanto afirmar que los ingresos laborales entre hombres y mujeres son diferentes.

Para contrastar la hipótesis de si los hombres ganan más que las mujeres, se modifica el código de la siguiente manera:

```
base_descriptiva %>%
  with(t.test(ingocup2~sex, alternative="greater"))

##
## Welch Two Sample t-test
##
## data: ingocup2 by sex
## t = 58.58, df = 133434, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1855.497      Inf
## sample estimates:
## mean in group 1 mean in group 2
##           7246.560          5337.457
```

Esto está asociado a las hipótesis:

$$H_o: \mu_1 \leq \mu_2$$

$$H_a: \mu_1 > \mu_2$$

Con esta prueba, rechazamos la hipótesis de que los hombres ganen menos o igual que las mujeres, puesto que el valor-*p* es muy pequeño. Por tanto, hay evidencia de que los ingresos de los hombres superan a los de las mujeres en las poblaciones. Por otro lado, vemos que se ha calculado también un intervalo unilateral; éste señala que, con 95% de confianza, los hombres ganan al menos 1 855.50 pesos más que las mujeres.

Estos ejercicios muestran la disparidad de los ingresos por la condición de ser hombre o mujer en México. No obstante, cuando comparamos los ingresos entre hombres y mujeres, debemos tener claro que hay jornadas de trabajo muy diferenciadas. Debido a que las mujeres tienen una sobrecarga de trabajo no remunerada, trabajan menos horas de manera remunerada fuera de la unidad doméstica. Por lo tanto, para tratar de aislar el efecto de ser hombre o mujer en los ingresos, muchas veces se recomienda comparar los ingresos por hora y no así los ingresos totales. Ello evidencia con más robustez estas diferencias.

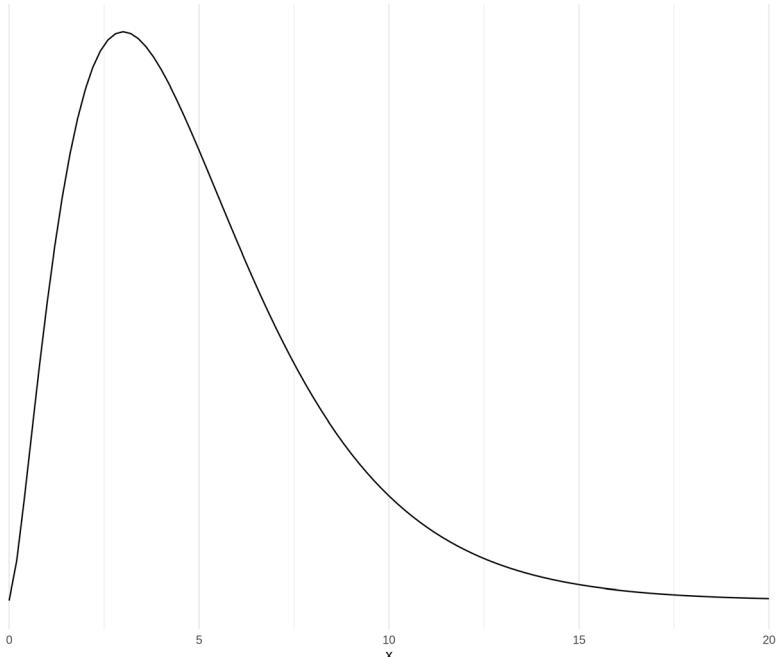
### C. Estimación de varianzas y sus pruebas de hipótesis

Así como podemos estimar las medias de la población, también podemos hacer inferencias acerca de las varianzas. Las varianzas, a diferencia de las medias, se distribuyen con la función chi-cuadrado. La distribución de chi-cuadrado ( $\chi^2$ ), con  $k$  grados de libertad, es la distribución de una suma de los cuadrados de  $k$  variables aleatorias distribuidas como normales estándar independientes. Esta distribución puede describir los comportamientos de las varianzas, puesto que una varianza, en su numerador, tiene una suma del cuadrado de las desviaciones. Al igual que la distribución *t*-Student, esta distribución tiene un parámetro de grados de libertad.

Una distribución chi-cuadrado se ve de esta forma:

```
ggchi <- ggplot(data.frame(x = c(0, 20)), aes(x = x)) +  
  stat_function(fun = dchisq, args = list(df = 5)) + ylab("") +  
  scale_y_continuous(breaks = NULL)  
ggchi + theme_minimal()
```

**Gráfica v-4. Distribución chi-cuadrado**



En este caso, no se trata de una distribución simétrica y es estrictamente positiva (los cuadrados de un valor nunca serán negativos).

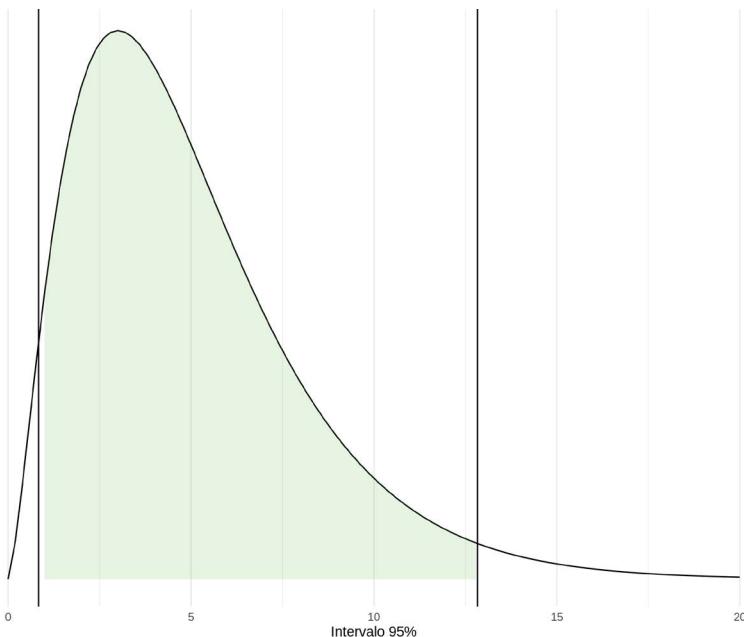
Esto es importante porque ya nuestros intervalos no provienen de sumar o restar un valor puesto que la distribución no es simétrica. Nuestro parámetro queda entre dos valores, pero centrado en el intervalo.

Para la varianza y su intervalo de confianza a nivel de  $(1 - \alpha) * 100\%$  se tiene que:

$$\frac{(n - 1)s^2}{\chi_{1-\alpha/2}^2} < \sigma^2 < \frac{(n - 1)s^2}{\chi_{\alpha/2}^2}$$

En la fórmula anterior,  $n$  se refiere al número de observaciones. Los valores de  $\chi^2$  tiene grados de libertad iguales a  $n - 1$  y están asociados al 95% central;  $s_2$  es la varianza muestral.

**Gráfica v-5. Distribución chi-cuadrado, con un área debajo de la curva de 0.95 central**



La inferencia sobre la varianza se puede hacer utilizando el comando “varTest()” del paquete *DescTools*.

```
base_descriptiva %>%
  with(VarTest(ingocup2))

##
## One Sample Chi-Square test on variance
##
## data: ingocup2
## X-squared = 5.1872e+12, df = 134510, p-value < 2.2e-16
## alternative hypothesis: true variance is not equal to 1
## 95 percent confidence interval:
## 38273851 38856767
## sample estimates:
## variance of x
##      38563649
```

Podemos también decir algo sobre el valor objetivo de nuestra hipótesis. Supongamos que quisiéramos contrastar contra la prueba de que la desviación estándar en la población es de 100 pesos. La varianza sería el cuadrado de este valor:

```
base_descriptiva %>%
  with(VarTest(ingocup2, sigma.squared = 100*100))

##
##  One Sample Chi-Square test on variance
##
##  data:  ingocup2
##  X-squared = 518719649, df = 134510, p-value < 2.2e-16
##  alternative hypothesis: true variance is not equal to 10000
##  95 percent confidence interval:
##  38273851 38856767
##  sample estimates:
##  variance of x
##            38563649
```

Se guardan como objeto nuestros resultados, lo que es siempre muy conveniente para pedir después o para realizar operaciones con ellos.

```
test2<-base_descriptiva %>%
  with(VarTest(ingocup2))
test2$conf.int

## [1] 38273851 38856767
## attr(,"conf.level")
## [1] 0.95

sqrt(test2$conf.int)

## [1] 6186.586 6233.520
## attr(,"conf.level")
## [1] 0.95
```

De esta manera, es un poco más sencillo interpretar los resultados puesto que podemos decir que, con 95% de confianza, la desviación estándar de los ingresos de los trabajadores mexicanos, para el periodo estudiado, se encuentra entre 6 186.59 y 6 233.52 pesos. Esto complementa las estimaciones que ya habíamos realizado y muestra que los trabajadores se alejan un valor considerable de la media.

Recuerda que con *tidy* esto también se nombra como una *tibble*:

```
tidy(test2)

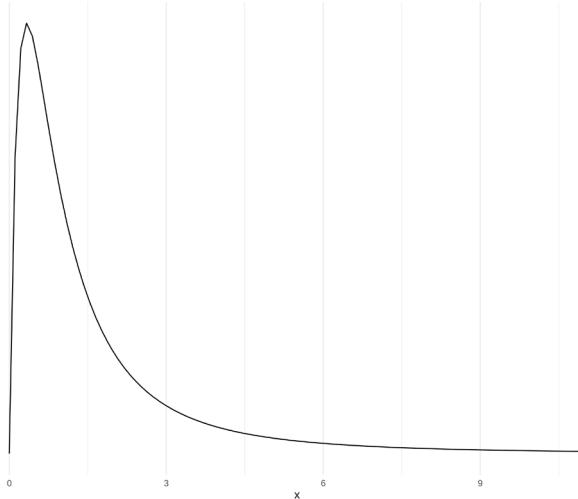
## # A tibble: 1 x 8
##   estimate statistic p.value parameter conf.low conf.high method
##   <dbl>     <dbl>    <dbl>    <dbl>      <dbl>    <chr>
## 1 3.86e7  5.19e12     0  134510  3.83e7 38856767. One S...
## # ... with 1 more variable: alternative <chr>
```

## D. Estimación de diferencias de varianzas y sus pruebas de hipótesis

La varianza muchas veces se puede asociar a la heterogeneidad de una población. Y la heterogeneidad de una variable puede indicar elementos sustantivos para el análisis. En nuestro ejemplo de los ingresos por trabajo, heterogeneidades más altas pueden estar asociadas a la desigualdad. Para comparar varianza, usamos su ratio; esto nos da un estadístico de prueba  $F$  para comparar dos muestras de poblaciones normales. Una razón de varianzas se distribuye como la función  $F$  de Fisher y Snedecor. Ésta se ve de la siguiente manera:

```
ggF <- ggplot(data.frame(x = c(0,11)), aes(x = x)) +
  stat_function(fun = df, args = list(df1 = 4, df2=4)) + ylab("") +
  scale_y_continuous(breaks = NULL)
ggF + theme_minimal()
```

**Gráfica v-6. Distribución  $F$**



Al igual que la  $\chi^2$ , esta distribución no es simétrica, pero está centrada en 1, puesto que es el valor que obtenemos cuando el numerador y denominador son iguales. Al provenir de dos variables que se distribuyen como  $\chi^2$ , habrá grados de libertad tanto en el numerador como en el denominador. Es decir,  $F$ :

$$\frac{U_1/gl_1}{U_2/gl_2} \sim F(gl_1, gl_2)$$

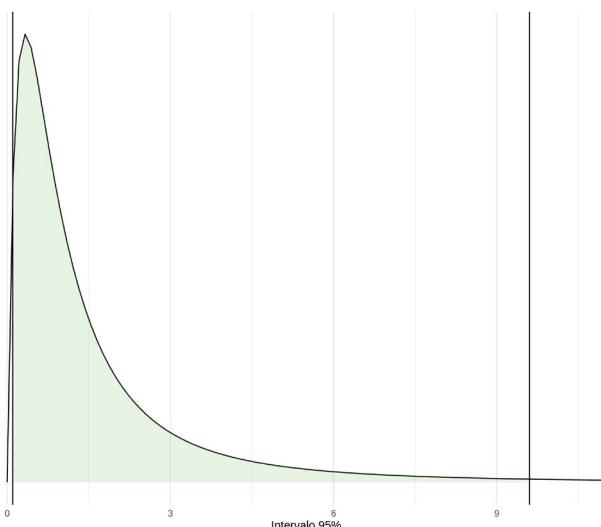
Donde  $U_1$  y  $U_2$  son variables independientes entre sí que se distribuyen como una  $\chi^2$  y que por tanto representan una suma de cuadrados, con  $gl_1$  y  $gl_2$  grados de libertad.

Para la estimación por intervalo, tenemos que se establece de la siguiente forma:

$$\frac{s_1^2}{s_2^2} \frac{1}{F_{gl_1, gl_2}} < \frac{\sigma_1^2}{\sigma_1^2} < \frac{s_1^2}{s_2^2} F_{gl_2, gl_1}$$

Donde  $s_1^2$  y  $s_2^2$  son las varianzas muestrales de cada grupo; mientras que los grados de libertad corresponden a los tamaños de las muestras menos 1, es decir,  $gl_1 = n_1 - 1$  y  $gl_2 = n_2 - 1$ . Los valores de  $F$  corresponden con los valores asociados donde se establece el 95% central de la distribución de probabilidad, para un intervalo de 95%, como se indica en la Gráfica v-7.

**Gráfica v-7. Distribución  $F$ , con un área debajo de la curva de 0.95 central**



Para comparar la varianza entre dos grupos usamos la función “var.test” de base de *R*, con el signo  $\sim$  para indicar la variable de dos grupos, y nos aseguramos que estamos comparando contra el ratio unitario que implica la igualdad de varianzas:

```
base_descriptiva %>%
  with(var.test(ingocup2~sex, ratio=1))

##
## F test to compare two variances
##
## data: ingocup2 by sex
## F = 1.6466, num df = 78556, denom df = 55953, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 1.621530 1.672028
## sample estimates:
## ratio of variances
## 1.646602
```

En el numerador queda la primera categoría (en este caso “hombres”) y en el denominador la segunda categoría. Es decir, el ratio de 1.65 indica que la varianza de los ingresos de los hombres es más grande que la de las mujeres. Los valores superiores a 1 implican que el valor del numerador es mayor; cuanto difiera de 1, tantas veces más grande será. Los valores inferiores a 1 indican que el denominador es mayor y cuánto 1 difiera de este valor, tanto más pequeño será. No es posible tener valores menores a 0.

Por otro lado, esta prueba nos dice que, con 95% de confiabilidad, la varianza de los hombres es 62.15% veces la varianza de las mujeres. Por lo que podríamos afirmar que los ingresos por trabajo de los varones son más heterogéneos que los ingresos de las mujeres.

En el caso de la prueba de hipótesis, la prueba realizada sería de este tipo:

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_a: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

Es decir, al obtener un valor-*p* muy pequeño, rechazamos la hipótesis nula. Por tanto, hay evidencia que las varianzas de los ingresos entre hombres y mujeres son diferentes.

En estas secciones observamos los elementos básicos de cuatro distribuciones que usaremos mucho en los procesos de inferencia: la distribución normal, la *t*-Student, la  $\chi^2$  y la *F*. A continuación revisaremos dos pruebas muy utilizadas en la investigación social.

## **E. Prueba chi-cuadrado: una aplicación para inferencia en tablas de contingencia**

Cuando tenemos dos variables cualitativas o nominales podemos hacer la prueba chi-cuadrado o prueba de independencia. Ésta tiene una lógica un poco diferente a las pruebas que hemos hecho hasta el momento porque proviene de comparar la distribución de los datos, dado que no hay independencia entre las variables y los datos que se tienen.

La hipótesis nula postula una distribución de probabilidad totalmente especificada, como el modelo matemático de la población que ha generado la muestra, por lo que, si la rechazamos, habrá evidencia estadística de la dependencia de las dos variables. Por ejemplo, para saber si hay dependencia entre la condición de ser hombre y mujer y la participación económica, podemos comparar los resultados del tabulado con lo que teóricamente esperamos si se mantienen la proporciones en cada una de las variables respecto a las categorías de la otra variable.

Recordemos el procedimiento para hacer una tabla de frecuencias de nuestras variables y revisemos cómo se distribuye la población en edad de trabajar:

```
tabla1<-base_descriptiva %>%
  filter(edad>14 & edad<99) %>% #filtra por población edad a trabajar
  mutate_at(vars(clase1, sex), as_label) %>% #etiqueta las variables
  tabyl(clase1, sex, show_missing_levels=F ) %>% # tabulado
  adorn_totals("row") %>% #añade una fila de total
  adorn_totals("col") #añade la columna de total

tabla1
##           clase1 Hombre Mujer Total
##   Población económicamente activa 109915 75087 185002
##   Población no económicamente activa 32906 83436 116342
##                               Total 142821 158523 301344
```

El recuento esperado viene del producto del total de la fila y del total de la columna, lo cual queda representado en la siguiente fórmula:

$$e_{ij} = \frac{\sum o_i * \sum o_j}{\sum o_{ij}}$$

Donde  $e_{ij}$  es el valor esperado de la celda correspondiente en la fila  $i$  y de la columna  $j$ ;  $o_i$  son los valores observados de cada fila  $i$ ;  $o_j$  es el valor observado de la columna  $j$ ; mientras  $o_{ij}$  son los valores observados en todas las filas y columnas de la tabla. Donde  $i$  llega hasta el número de filas  $r$  y  $j$  hasta el número de columnas  $c$ .

Por ejemplo, en nuestra tabla lo esperado es que la celda (1,1),  $e_{1,1}$  se viera así:

$$e_{1,1} = \frac{185002 * 142821}{301344} = 87681.09$$

Este valor es menor al que estimamos en el tabulado. Así todas las celdas pueden diferir de nuestro valor esperado. Una vez calculados los valores para todas las celdas, podemos calcular una medida. El estadístico de prueba  $\chi^2$  tiene la siguiente fórmula:

$$\chi^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Como observamos, este estadístico es la suma de un cuadrado; por tanto, se distribuye como una chi-cuadrada cuyos grados de libertad están dados por  $r - 1 * c - 1$ . Para hacer esta prueba con el código que da la tabla, añadimos la función “chisq.test()”, proveniente del mismo paquete *janitor*, que nos provee la función “tabyl()”:

```
base_descriptiva %>%
  filter(edad>14 & edad<99) %>% #filtra por población edad a trabajar
  mutate_at(vars(clase1, sex), as_label) %>% #etiqueta las variables
  tabyl(clase1, sex, show_missing_levels=F ) %>% # tabulado
  chisq.test()

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: .
## X-squared = 27759, df = 1, p-value < 2.2e-16
```

Para este caso en particular, lo anterior se relaciona con las siguientes hipótesis:

$H_o$ : La condición de hombre o mujer y la participación económica son independientes.

$H_a$ : La condición de hombre o mujer y la participación económica no son independientes.

De ahí que rechacemos la  $H_o$  y podamos afirmar  $H_a$ ; las variables no son independientes dado que el valor-*p* reportado es muy pequeño y por debajo de 0.05 de significancia. Es importante señalar que esta prueba no nos dice el sentido, pero lo que se observa en la tabla da cuenta de que los hombres tienen mayor participación económica que las mujeres.

## F. Análisis de la varianza de un factor: comparación de varias medias

El análisis de varianza (ANOVA, por sus siglas en inglés “Analysis of Variance”) es un modelo estadístico. Los modelos tienen supuestos que hay que respetar, como toda simplificación de la realidad. En particular, esta herramienta compara dos varianzas: la explicada por el modelo y la no explicada. Al comparar varianzas, nos remite por completo a la distribución *F* que ya revisamos en secciones anteriores.

El análisis de la varianza compara la variación debida a unas determinadas fuentes con la variación existente entre individuos que deberían ser similares. En particular, la prueba ANOVA contrasta si varias poblaciones tienen la misma media, comparando lo separadas que están entre sí las medias muestrales en relación con la variación existente dentro de las muestras (Moore, 2010:661).

Para este texto nos quedaremos con el análisis de varianza más simple, cuyo objetivo es establecer el efecto de un factor (una variable con varias categorías) sobre una variable cuantitativa, conocido como simple o *oneway*. Este tipo de aplicación también se puede entender como la comparación de muestras de una variable cualitativa de más de dos categorías. En nuestra aplicación con la ENOE, revisaremos si la región de residencia de los trabajadores tiene un efecto en la distribución de los ingresos por trabajo.

## a. Apreciación gráfica

Uno de los elementos fundamentales es que nuestros grupos observados deben provenir de un muestreo aleatorio y ser independientes. Pero también, como veremos más adelante, la ANOVA se basa en que nuestra variable cuantitativa es normal. Como sabemos por la revisión del capítulo anterior, los ingresos no son normales, por lo que vamos a hacer una transformación de nuestra variable en estudio. En este caso, trabajaremos con los logaritmos. Del mismo modo, nuestra variable cualitativa tiene etiquetas muy largas, por lo que las vamos a modificar.

Con el siguiente código cambiamos las etiquetas:

```
base_descriptiva$t_loc<-set_labels(base_descriptiva$t_loc,
                                     labels=c("100,000 hab. +",
                                             "15,000-99,999",
                                             "2,500 a 14,999",
                                             "2,500 y menos"))
```

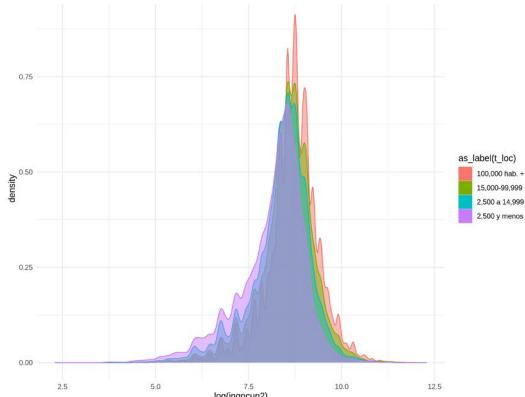
Y con el siguiente hacemos una gráfica de densidad de la variable de ingresos laborales según el tamaño de localidad de la residencia:

```
g_anova <-base_descriptiva %>%
  ggplot(aes(x=log(ingocup2), fill=as_label(t_loc),
             color=as_label(t_loc),
             alpha=I(0.5)))

g_anova + geom_density() + theme_minimal()

## Warning: Removed 271080 rows containing non-finite values (stat_density).
```

**Gráfica v-8. Distribución del logaritmo de los ingresos según tamaño de la localidad de la residencia de los trabajadores.  
México, trimestre III, 2019**



FUENTE: elaboración propia con base en la ENOE.

Este primer acercamiento nos permite ver algunas diferencias en la distribución. Las localidades urbanizadas tienen mayores ingresos y las menos urbanizadas tienen menos. ¿Cómo podemos verificar estos acercamientos gráficos y saber si lo que observamos en la muestra se mantiene en la población? Al incluir el componente aleatorio, y por tanto la probabilidad, podemos inferir sobre las poblaciones.

Por ello, debemos establecer las pruebas de hipótesis correspondientes. En términos de comparación de medias, éstas se escribirían de la siguiente manera:

$$H_o: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_a: \text{Alguna de las medias es diferente}$$

Este tipo de pruebas es global puesto que, con dos de las medias de las cuatro que estemos estudiando, la prueba resultará en que podamos rechazar la hipótesis nula. Por otro lado, en términos de varianzas las hipótesis se pueden escribir de la siguiente forma:

$$H_o: \frac{\text{Variación entre las medias muestrales}}{\text{Variación entre individuos de la misma muestra}} \leq 1$$

$$H_a: \frac{\text{Variación entre las medias muestrales}}{\text{Variación entre individuos de la misma muestra}} > 1$$

Es decir, comparamos la varianza que proviene, en nuestro caso particular, de la residencia según el tamaño por localidad; o por otra razón. Supongamos que decimos que los ingresos están explicados por la residencia, entonces dos personas que residen en un mismo tipo de localidad deberían tener ingresos muy parecidos; si no, los ingresos están siendo explicados por otros motivos. De ahí que comparemos ambas varianzas a partir de un estadístico  $F$  que se contrasta contra el valor de 1, de igualdad.

La prueba es unilateral de lado derecho, pues lo que queremos contrastar es que la varianza de las medias entre las localidades es mayor a lo no explicado por esta residencia. Ello implica que el factor analizado (residencia según tamaño de localidad) explica la varianza de la variable cuantitativa en análisis (ingresos laborales mensuales).

La prueba ANOVA, o análisis de varianza en *R*, se lleva a cabo con el comando “*aov()*”, así como de los comandos de “*stats*” que están predeterminados en la instalación del programa. Se ejecutaría de la siguiente forma:

```
anova<-base_descriptiva %>%
  filter(ingocup2>0) %>% #Eliminamos Los cero para no tener valores no definidos
al sacar logaritmo
  with(aov(log(ingocup2) ~ as_label(t_loc),
na.action=na.omit))

tidy(anova)
## # A tibble: 2 × 6
##   term            df    sumsq   meansq statistic p.value
##   <chr>           <dbl>  <dbl>    <dbl>     <dbl>    <dbl>
## 1 as_label(t_loc)     3  5440.  1813.     3136.     0
## 2 Residuals        126516 73153.    0.578     NA      NA
```

Estos valores nos indican que nuestro estadístico *F* es igual a 3135.85, siendo mayor que 1, lo que asocia un valor-*p* muy pequeño. Por tanto, rechazamos la  $H_0$  y podemos afirmar que al menos una de las medias muestrales de los ingresos, según el tamaño de la localidad de residencia, son diferentes; o bien, que nuestro factor de residencia explica la varianza de nuestros ingresos laborales.

En la tabla de resultados, podemos observar que la variación de las medias muestrales es 1813.19 y la variación entre los individuos de la muestra es de 0.58; el ratio de estos valores establece el valor de nuestro estadístico *F*.

## b. Comparación entre grupos

Al ser una prueba global, tenemos información que —en efecto— los ingresos laborales difieren según la residencia por tamaño de localidad. Pero, ¿cuáles son esas diferencias entre los grupos?

En *R* podemos utilizar la función “*TukeyHSD()*”, que proviene de la abreviación de *Tukey's 'Honest Significant Difference' method*. Esta función hará los pares posibles y brindará el intervalo de las estimaciones de sus diferencias, así como el valor-*p* asociado a la prueba de hipótesis de su igualdad bivariada. Puedes revisar más en la ayuda de este comando:

TukeyHSD(anova)

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = log(ingocup2) ~ as_label(t_loc), na.action = na.omit)
##
## $`as_label(t_loc)`
##          diff      lwr      upr p adj
## 15,000-99,999-100,000 hab. + -0.1702642 -0.1867656 -0.1537629 0
## 2,500 a 14,999-100,000 hab. + -0.3368634 -0.3539275 -0.3197993 0
## 2,500 y menos-100,000 hab. + -0.5825655 -0.5991671 -0.5659640 0
## 2,500 a 14,999-15,000-99,999 -0.1665992 -0.1881380 -0.1450603 0
## 2,500 y menos-15,000-99,999 -0.4123013 -0.4334756 -0.3911270 0
## 2,500 y menos-2,500 a 14,999 -0.2457021 -0.2673179 -0.2240864 0
```

De acuerdo con los resultados, todos los valores-*p* son muy pequeños; de ahí que todas las diferencias entre pares podrían establecerse para las poblaciones. Por esto podemos afirmar que la residencia tiene un efecto para explicar los ingresos. Además, las medias de los logaritmos de los ingresos entre todas las comparaciones posibles son diferentes entre sí. Aunque, antes de afirmar esto con tanta certeza, debemos evaluar si cumplimos con los supuestos del modelo que estamos realizando.

### c. Supuestos del análisis de varianza

Son tres:

- Las observaciones se obtienen de forma independiente y aleatoria de la población definida por los niveles del factor.
- Los datos de cada nivel de factor se distribuyen normalmente.
- Estas poblaciones normales tienen una varianza común.

El primer supuesto proviene del diseño muestral; mientras que los otros dos los podemos verificar con otras pruebas de hipótesis. A continuación mostramos la prueba para cada uno, con sus hipótesis nula y alternativa, teniendo en cuenta que no siempre conviene rechazar la  $H_0$ .

#### i) Prueba de normalidad

Si bien ya hicimos una gráfica, pudimos notar que aún con la transformación de logaritmos de los ingresos ésta distaba mucho de parecerse a una normal.

```
base_descriptiva %>%
  filter(ingocup2>0) %>% #filtra los ceros por no tener logaritmos indefinidos
  with(ks.test(log(ingocup2), #la variable a revisar su normalidad
               "pnorm", #Aquí seleccionamos que revisamos la distribución en
               comparación a una normal
               mean=mean(log(ingocup2)), # La media
               sd=sd(log(ingocup2)))) # La desviación estándar

## Warning in ks.test(log(ingocup2), "pnorm", mean = mean(log(ingocup2)), sd
## = sd(log(ingocup2))): ties should not be present for the Kolmogorov-Smirnov
## test

##
## One-sample Kolmogorov-Smirnov test
##
## data: log(ingocup2)
## D = 0.11043, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Las hipótesis son:

$H_o$ : Los ingresos laborales se distribuyen como una normal

$H_a$ : Los ingresos laborales no se distribuyen como una normal

Según el valor- $p$  pequeño de la prueba, menor a 0.05, rechazamos la hipótesis nula y, por tanto, los ingresos laborales; según la evidencia estadística, no se comportan normalmente. Ello quiere decir que no cumplimos con el supuesto de normalidad.

### *ii) Prueba de Bartlett para homogeneidad de varianzas*

Esta prueba compara las varianzas de más de dos grupos.

```
base_descriptiva %>%
  filter(ingocup2>0) %>% #Eliminamos los cero para no tener valores no definidos
  al sacar logaritmo
  with(bartlett.test(log(ingocup2) ~ as_label(t_loc)))

##
## Bartlett test of homogeneity of variances
##
## data: log(ingocup2) by as_label(t_loc)
## Bartlett's K-squared = 1801.9, df = 3, p-value < 2.2e-16
```

$H_o$ : Las varianzas entre los grupos son iguales

$H_a$ : Las varianzas entre los grupos son diferentes

De nuevo, el valor-*p* es muy pequeño. Lo que implica que rechazamos la hipótesis nula de homogeneidad y, por tanto, hay evidencia de heterogeneidad. En consecuencia, tampoco se cumple este supuesto.

#### d. ¿Qué hacer? Un método: la prueba *Kruskal-Wallis test*

No cumplir supuestos es parte normal del análisis estadístico; ello nos lleva a buscar y aplicar técnicas más robustas. En este caso, una técnica más robusta (pero no la única) es aplicar la prueba de Kruskal-Wallis. Esta prueba es muy parecida, pero se basa en el orden de las observaciones y no en las variables originales. Para interpretarse, se lee muy parecida a la ANOVA: “[...] la prueba de Kruskal-Wallis de rangos se basa en ordenar las respuestas de todos los grupos, y a continuación aplicar el ANOVA de un factor a los rangos y no a los valores originales” (Moore, 2010, p. 768).

Para aplicarla en *R*, se utiliza el comando “kruskal.test()”.

```
kruskal<-base_descriptiva %>%  
  filter(ingocup>0) %% #filtrar Los casos 0  
  with(kruskal.test(log(ingocup2) ~ as_label(t_loc))) #aplica la prueba  
kruskal # imprime el objeto  
  
##  
## Kruskal-Wallis rank sum test  
##  
## data: log(ingocup2) by as_label(t_loc)  
## Kruskal-Wallis chi-squared = 8199, df = 3, p-value < 2.2e-16  
  
tidy(kruskal) # imprime en formato tabular  
  
## # A tibble: 1 x 4  
##   statistic p.value parameter method  
##     <dbl>    <dbl>      <int> <chr>  
## 1      8199.     0          3 Kruskal-Wallis rank sum test
```

$H_o$ : Las poblaciones tienen la misma distribución

$H_a$ : Las poblaciones no tienen la misma distribución

Es decir, rechazamos la  $H_o$  y decimos que las poblaciones no provienen de una misma distribución; por lo tanto, son diferentes.

Para ver las comparaciones, tenemos que usar el “Dunntest()”, del paquete *DescTools*, y ello nos revela una tabla muy parecida al “Tukey HSD”:

```
base_descriptiva %>%
  filter(ingocup2 > 0) %>%
  with(DunnTest(log(ingocup2) ~ as_label(t_loc)))

##
## Dunn's test of multiple comparisons using rank sums : holm
##
##          mean.rank.diff   pval
## 15,000-99,999-100,000 hab. +    -8678.755 <2e-16 ***
## 2,500 a 14,999-100,000 hab. +   -15892.124 <2e-16 ***
## 2,500 y menos-100,000 hab. +   -25671.216 <2e-16 ***
## 2,500 a 14,999-15,000-99,999   -7213.370 <2e-16 ***
## 2,500 y menos-15,000-99,999   -16992.461 <2e-16 ***
## 2,500 y menos-2,500 a 14,999   -9779.091 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En este capítulo se planteó un breve recorrido por las estimaciones por intervalo de medias, varianzas muestrales y la estimación de la diferencia de éstas entre dos grupos. Al mismo tiempo, se presentaron dos pruebas de hipótesis básicas como la  $\chi^2$  para comprender la inferencia estadística. Nos acercamos a los modelos a partir del análisis de la varianza de un factor sobre una variable cuantitativa.

En este capítulo dimos cuenta de cómo el proceso de evidencia estadística incluye revisar supuestos y tomar decisiones sobre qué técnicas utilizar. Esto es un proceso al que todo investigador se enfrenta. Los datos no son los equivocados, más bien, debemos encontrar las técnicas que nos permitan evidenciar con más robustez nuestros resultados.

Con respecto al mercado laboral mexicano, continuamos aportando evidencia de las desigualdades. En este caso se ha establecido, con evidencia para la población, que los hombres tienen ingresos superiores al de las mujeres; mientras que las disparidades territoriales también fueron evidenciadas. Las disparidades más grandes suceden entre los estratos más urbanizados y los más rurales.

# **VI. La regresión lineal. Un ejemplo para modelar los ingresos en los mercados laborales**

## **Introducción**

En el Capítulo IV aprendimos sobre la relación entre dos variables cuantitativas: entre los ingresos laborales, los cuales ya habíamos identificado como válidos, y los años de escolaridad.

Esta relación ha sido ampliamente estudiada desde la economía laboral. Fue propuesta por Mincer (1974) y, si bien, se ha reformulado y se han hecho acotaciones muy importantes sobre cómo se deben medir de manera más certera los retornos de la educación (Heckman, Lochner & Todd, 2003, 2006), parece ser que ésta es una de las variables que más nos permite explicar la varianza de una variable tan heterogénea como los ingresos laborales en México.

En el capítulo IV revisamos de manera descriptiva e introductoria esta relación. Si bien la correlación mide la fuerza y la dirección de la relación lineal entre estas dos variables, complementamos la información con un diagrama de dispersión que permite visibilizarla (puede ser lineal o no). En este capítulo vamos a establecer un modelo que dibuje de manera precisa una línea entre la nube de puntos de las observaciones. En específico, aprenderemos a dibujar una línea recta para la muestra y estableceremos si esta línea que describe la muestra también es válida en términos poblacionales, aplicando lo visto en el capítulo V en términos de inferencia.

Nuevamente, recomendamos leer los libros de estadística sobre el tema. Aquí sólo pondremos los elementos mínimos; no obstante, hay libros muy interesantes sobre la regresión lineal y otras extensiones (Angrist & Pischke, 2009; Wooldridge, 2010).

Para revisar los elementos básicos de la regresión lineal hemos organizado el capítulo como sigue. Primero, introducimos a la regresión por medio de mínimos cuadrados ordinarios (MCO). Luego,

revisamos el modelo más básico de la regresión lineal simple. En una tercera sección, discutimos sobre cómo complejizar el análisis con más variables explicativas. La cuarta parte provee los códigos para la presentación de resultados. En la quinta parte, se menciona un elemento importante, el cual refiere a la comparación de efectos a partir de variables estandarizadas. Finalmente, cerramos el capítulo con algunas opciones en caso de no cumplir con los supuestos del modelo.

Para comenzar, vamos a cargar los paquetes o instalarlos si es necesario. Recuerda que puedes descargar el ambiente desde <<https://tinyurl.com/Base-Enoe>>.

```
if (!require("pacman")) install.packages("pacman") # instala pacman si se requiere
## Loading required package: pacman
pacman::p_load(tidyverse, haven, sjlabelled, janitor, broom, DescTools, car, sjPlot,
stargazer, lm.beta, Rcurl, robustbase) #carga Los paquetes necesarios para este capítulo
load("base_descriptiva2.RData") ## recuerda tenerla descargada en tu carpeta de trabajo
```

## A. La línea de mínimos cuadrados ordinarios “(MCO)”

Si imaginamos una línea entre los puntos de dispersión, buscaremos aquella que minimice la distancia entre los puntos; a esta línea se le conoce como línea de mínimos cuadrados ordinarios.

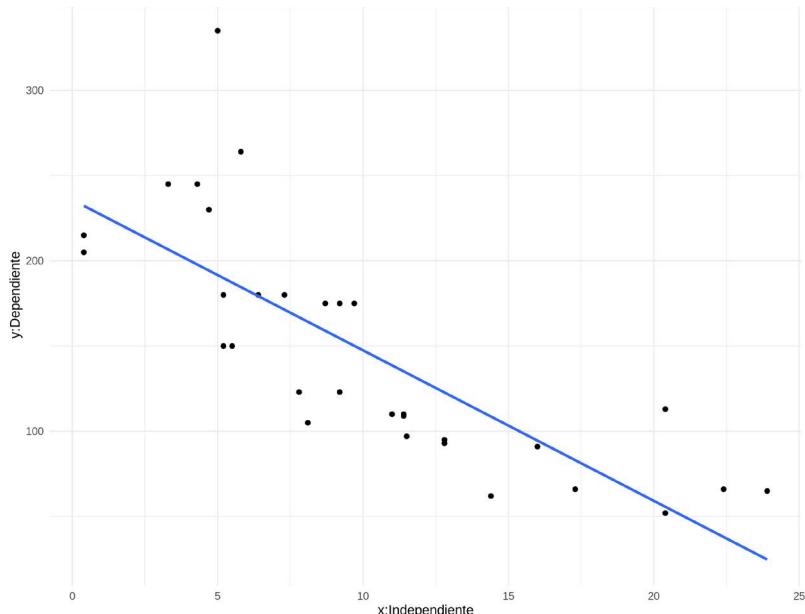
En este caso, tenemos la claridad teórica de que una variable,  $y$ , está siendo explicada por otra,  $x$ . De ahí que, en nuestra gráfica, la variable dependiente, de respuesta, predicha o regresando se coloca en el eje vertical. La variable independiente, de control, predictora o regresor se coloca en el eje horizontal. De tal cuenta que nuestra relación se aprecia en la Gráfica VI-1 en página siguiente.

Como se ve, la línea tiene una inclinación —en este caso negativa— que llamamos pendiente. La pendiente también puede ser entendida como la tasa de cambio; es decir, lo que cambia  $y$ , dado un cambio en  $x$ . También la línea tiene otro valor importante: un intercepto; es decir, el valor de  $y$ , cuando  $x = 0$ .

Si estamos haciendo un análisis descriptivo, muestral, podemos encontrar la pendiente y la intersección de la línea de mejor ajuste utilizando la fórmula:

$$y = b_0 + b_1x$$

**Gráfica vi-1. Ejemplo de una gráfica de dispersión con la línea ajustada por MCO**



Para encontrar los valores de  $b_0$  y  $b_1$ , tenemos una fórmula para cada uno de estos estadísticos. Estos provienen de un proceso de minimización.

$$b_1 = r \frac{s_y}{s_x}$$

Donde  $r$  es el coeficiente de correlación, y  $s_y$  y  $s_x$  son las desviaciones estándar de las variables en estudio.

$$b_0 = y - b_1 \bar{x}$$

Para la población, como hemos observado, tenemos que usar letras griegas y también incluir un elemento de error para nuestras estimaciones. De ahí que las anteriores ecuaciones se escriben en términos poblacionales de la siguiente forma:

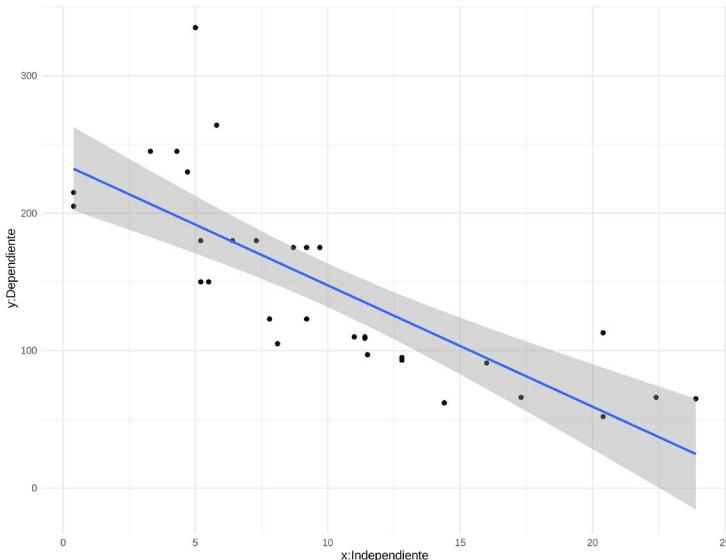
$$y = \beta_0 + \beta_1 x + \epsilon$$

Donde los parámetros  $\beta_0$  y  $\beta_1$  describen el intercepto y la pendiente de la población, respectivamente.

Como establecimos en el capítulo anterior, al hacer estimaciones poblacionales sabemos que tenemos el valor de un parámetro y un cierto margen que dependerán de los niveles de confiabilidad con la que queremos estimar. Por ejemplo, la Gráfica vi-2 es muy similar a la Gráfica vi-1, pero vemos una sombra gris. Esta sombra representa estimaciones con 95% de confianza.

Pensemos que los parámetros  $\beta_0$  y  $\beta_1$  tienen estimaciones que podemos encontrar con nuestra muestra y cierto margen. De ahí que, en esa estimación muestral marcada por la línea en azul, podemos saber que la línea para la población estará en algún lugar de la línea gris. En otras palabras, para trasladar la línea de MCO a la población, necesitamos estimar también intervalos de confianza, como se observa en la gráfica de acuerdo con la sombra gris.

**Gráfica vi-2. Ejemplo de una gráfica de dispersión con la línea ajustada por mco con intervalos de confianza**



Otro concepto importante a tener en cuenta son los residuos. Como observamos, no todos los puntos pasan por la recta; por lo que los valores predichos por la recta calculada no siempre corresponden al valor real de  $y$ . Lo predicho o estimado podemos denotarlo con un “gorro” (ácento circunflejo); de ahí que los residuos resultan de la diferencia entre lo predicho y lo estimado:  $\hat{y} - y$ . La suma de todos los residuos debería ser cercana a 0. Estos valores también pueden ser llamados errores o perturbaciones.

Estos elementos básicos nos permiten iniciar nuestra práctica, pero instamos a quien lee este libro a profundizar en otros textos de especialización estadística. Nos disponemos en las siguientes páginas a presentar un ejemplo práctico utilizando la ENOE.

## B. Regresión lineal simple

En este capítulo vamos a replicar un modelo muy básico. Buscamos explicar los ingresos por hora (válidos, sin valores perdidos y mayores que cero) según los años de escolaridad como variable explicativa. Luego agregaremos otras variables que serán controles de esta relación.

Para ello, vamos a establecer un subconjunto de nuestras posibles variables explicativas, que son variables con las cuales hicimos análisis anteriores. Hacer esta operación es importante porque sólo podemos comparar modelos en términos de sus indicadores de ajuste y predicción si tienen la misma cantidad de observaciones.

```
dmodelo<- base_descriptiva %>%
  filter(clase2==1 & anios_esc<99 & ingocup2>0 & eda<99 & eda>14) %>% ## casos
  válidos
  select(ingocup2, t_loc, eda, sex, anios_esc, fac) %>%
    na.omit() # elimina renglones si hay un valor perdido
tail(dmodelo) #muestra los últimos 6 casos

##   ingocup2 t_loc eda sex anios_esc fac
## 125624     2000    1  53   2      9 266
## 125625     6450    1  40   1      6 266
## 125626     3440    1  20   1      9 266
## 125627     3440    1  18   1      9 266
## 125628     6020    1  57   2      6 266
## 125629    12900    1  43   1      9 266
```

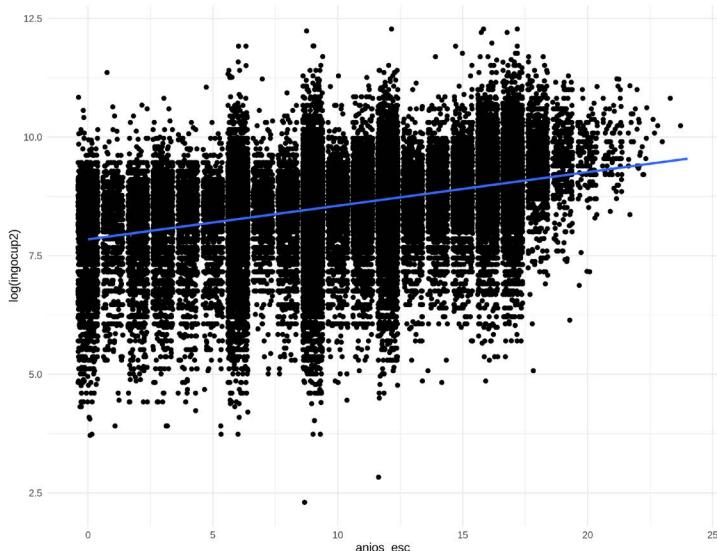
Primero vamos a establecer la parte descriptiva de la relación que queremos plantear entre los ingresos y los años de escolaridad que, como ya se dijo, es una relación bien establecida en la literatura

de mercados de trabajo. Esta relación se ha mostrado también en estudios recientes para México, aunque se ha establecido que debe ser estudiada en términos de la evolución de los rendimientos, de la cohorte, de los sectores y de los niveles de sobreeducación (Binelli & Rubio, 2013; Cendejas, 2017; Escoto, Márquez & Prieto, 2020; Levy & López, 2019; Levy & Székely, 2016; Linthon, 2018; Caamal, 2017).

Iniciemos nuestro análisis replicando la gráfica de dispersión realizada con *ggplot2*.<sup>14</sup>

```
gg <- dmodelo %>%
  ggplot(aes(anios_esc, log(ingocup2))) # el Lienzo
gg + geom_jitter() + #la geometría jitter
  geom_smooth(method="lm") + # geometría que ajusta una Línea recta (lm=linear model
  theme_minimal() # cambia apariencia
```

**Gráfica vi-3. Diagrama de dispersión de la relación entre los ingresos y los años de escolaridad. México, trimestre III, 2019. Personas con remuneraciones válidas y mayores que cero**



FUENTE: elaboración propia con base en la ENOE.

<sup>14</sup> La diferencia de la Gráfica VI-3 con la Gráfica IV-12 es que en la VI-3 no se consideran los ceros. Es decir, el cálculo es con trabajadores remunerados.

Para calcular la correlación de estas variables, creamos la variable dentro de nuestro objeto.

```
dmodelo<-dmodelo %>%
  mutate(log_ingocup= log(ingocup2))
```

Además, podemos iniciar con una prueba de hipótesis sobre la correlación usando el paquete *broom* (“escoba”). Como señalamos anteriormente, es un paquete que limpia y transforma objetos con estimaciones de inferencias en elementos de tipo “tibble” o *dataframe*. Esto es útil si los queremos graficar después.

```
dmodelo %>%
  with(cor(log_ingocup, anios_esc, use = "pairwise")) # estimación muestral
## [1] 0.3804431

cor_test<-dmodelo %>%
  with(cor.test(log_ingocup, anios_esc, use = "pairwise")) # prueba de hipótesis.

#dos modos de visualizar el resultado
cor_test

##
## Pearson's product-moment correlation
##
## data: log_ingocup and anios_esc
## t = 145.81, df = 125627, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.3757037 0.3851625
## sample estimates:
## cor
## 0.3804431

tidy(cor_test)

## # A tibble: 1 x 8
##   estimate statistic p.value parameter conf.low conf.high method
##   <dbl>     <dbl>    <dbl>     <int>    <dbl>    <dbl>   <chr>
## 1  0.380     146.      0     125627   0.376    0.385 Pears...
## # ... with 1 more variable: alternative <chr>
```

Esta prueba de hipótesis no establece si la correlación poblacional es distinta de cero o no. Es decir, las hipótesis nula y alternativa de este test son:

$$H_o: \rho = 0$$

$$H_a: \rho \neq 0$$

Donde  $\rho$  es el coeficiente de Pearson de correlación poblacional. Es decir,  $r$  es nuestro estimador. Al haber un valor- $p$  muy pequeño, rechazamos la hipótesis, lo cual nos da información acerca de que la correlación en la población ( $\rho$ ) es distinta de cero.

### a. Aplicación de una regresión lineal simple

Como ya se vio, la regresión lineal nos ayuda a describir esta relación a través de una línea recta. Los estimadores de la recta para la muestra deben ser puestos a prueba, así como el ajuste del modelo global. De ahí que se hagan pruebas de hipótesis para los coeficientes denotados por la letra griega  $\beta$ ; mientras que el modelo total explicará la varianza de la variable dependiente y será importante el qué tanto.

El modelo de regresión nos dará estimadores de los parámetros que describen la recta de regresión y, además, nos dará información sobre estos estimadores en términos de intervalos de confianza y de pruebas de especificación.

Una vez transformada nuestra variable, corremos el modelo que nos da los coeficientes de la línea MCO:

```
modelo <- lm(log_ingocup ~ anios_esc, data=dmodelo)
summary(modelo) # resultado formal

##
## Call:
## lm(formula = log_ingocup ~ anios_esc, data = dmodelo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -6.1814 -0.3300  0.0754  0.4312  3.7536 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 7.8468148  0.0053398 1469.5   <2e-16 ***
## anios_esc   0.0707979  0.0004856   145.8   <2e-16 ***
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7232 on 125627 degrees of freedom
## Multiple R-squared:  0.1447, Adjusted R-squared:  0.1447 
## F-statistic: 2.126e+04 on 1 and 125627 DF,  p-value: < 2.2e-16
```

Esta tabla es bastante complicada, pero tiene la información necesaria. Revisaremos cada una de las partes del modelo, empezando con el ajuste de los coeficientes. Para obtenerlos, aplicamos el comando “tidy()”.

```
tidy(modelo) # pruebas de hipótesis de los coeficientes

## # A tibble: 2 x 5
##   term      estimate std.error statistic p.value
##   <chr>     <dbl>    <dbl>     <dbl>    <dbl>
## 1 (Intercept) 7.85    0.00534    1469.     0
## 2 anios_esc   0.0708   0.000486    146.     0
```

Al tener una escala logarítmica, la variable dependiente en términos del cambio en  $x$  se lee como un cambio porcentual.<sup>15</sup> Nuestros resultados indican que, ante un cambio de un año de escolaridad, los ingresos aumentarían 7.08%; siendo esta estimación diferente de cero para la población, según lo que se rechaza en la hipótesis nula. Como observamos en la tabla, los estimadores tienen una prueba de hipótesis que se trata de una prueba  $t$ . Las hipótesis que se establecen son como siguen:

$$H_o: \beta_0 = 0 \text{ vs. } H_a: \beta_0 \neq 0$$

$$H_o: \beta_1 = 0 \text{ vs. } H_a: \beta_1 \neq 0$$

La significancia estadística del intercepto es menos fundamental, puesto que nos interesa saber si la variable independiente  $x$  funciona para predecir  $y$ . En caso contrario, el valor de  $y$  no cambiaría, independientemente del valor de  $x$ . Esto implica que la pendiente de la recta,  $\beta_1$ , es cero. Estas pruebas, así como los intervalos de confianza de los estimadores, nos permiten hacer inferencia de esta capacidad de predicción para la población. Para obtener los intervalos de confianza, podemos hacerlo a partir del siguiente comando:

```
confint(modelo)

##             2.5 %    97.5 %
## (Intercept) 7.83634893 7.85728075
## anios_esc   0.06984627 0.07174963
```

En el caso de los intervalos, con 95% de confianza, podemos establecer que este cambio en la población está entre 6.98% y 7.17%.

---

<sup>15</sup> Se recomienda revisar con detalle el capítulo 2 de Wooldridge (2010) que establece cómo se leen las transformaciones logarítmicas.

Para el ajuste global del modelo, podemos utilizar el comando “`glance()`” sobre el objeto que creamos; ello nos dará la información correspondiente:

```
glance(modelo) # resultado ajuste global

## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic p.value    df logLik     AIC
##       <dbl>         <dbl> <dbl>      <dbl>   <dbl> <dbl> <dbl>
## 1     0.145        0.145  0.723    21260.      0     2 -1.38e5 2.75e5
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>
```

Otra manera de ver este ajuste es con el comando “`anova()`”:

```
anova(modelo)

## Analysis of Variance Table
##
## Response: log_ingocup
##             Df Sum Sq Mean Sq F value    Pr(>F)
## anios_esc     1 11120 11119.7 21260 < 2.2e-16 ***
## Residuals 125627 65707     0.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En efecto, esta información es sobre el análisis de varianza del modelo. Queremos analizar la varianza de la variable dependiente (ingresos laborales) explicada por la variable independiente, en este caso, años de escolaridad. Al comparar esta varianza tenemos una prueba  $F$ , que responde a las siguientes hipótesis:

$H_o$ : La varianza explicada por el modelo es igual a lo no explicada

$H_a$ : La varianza explicada por el modelo es mayor a lo no explicada

También tenemos la estimación del estadístico  $R^2$ , el cual se puede considerar como la proporción de la variación total en la variable de respuesta. En nuestro caso, los ingresos laborales se pueden explicar usando la variable independiente  $x$ , años de escolaridad, en el modelo. El modelo ajustado explica alrededor de 14.5% de la variancia total.

## b. Supuestos y su diagnóstico

Como señalamos en el capítulo anterior, los modelos de regresión tienen supuestos. Vamos a revisar los supuestos principales. Seguimos a Moore (2010: 709-710) en términos de la exposición de estos supuestos y usamos la librería *car* (*Companion to Applied Regression*) para su evaluación.

- *Supuesto 1:* Las observaciones son independientes. No se pueden tener observaciones repetidas de un mismo individuo.
- *Supuesto 2:* La verdadera relación, en la población, es lineal. No podemos observar la verdadera recta de regresión y es muy raro observar una perfecta relación lineal en nuestros datos.
- *Supuesto 3:* La desviación típica de la respuesta a lo largo de la verdadera recta es siempre la misma u homocedasticidad de los errores.
- *Supuesto 4:* La respuesta varía normalmente en relación con la verdadera recta de regresión.

Woolridge (2010) también añade el supuesto de aleatoriedad en las observaciones.

En el caso de los supuestos 1 y 2, establecemos que estos se cumplen. Nos vamos a concentrar en evaluar los supuestos 3 y 4. Para ello, tendremos elementos gráficos y pruebas de hipótesis.

Para evaluar la homocedasticidad, recurrimos a un test para verificar esta varianza constante de los errores:

```
# non-constant error variance test
ncvTest(modelo)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1471.015, Df = 1, p = < 2.22e-16
```

Como observamos, este test da un estadístico de prueba que se distribuye como una chi-cuadrada. Las hipótesis de esta prueba son:

$H_o$ : Los errores son homocedásticos

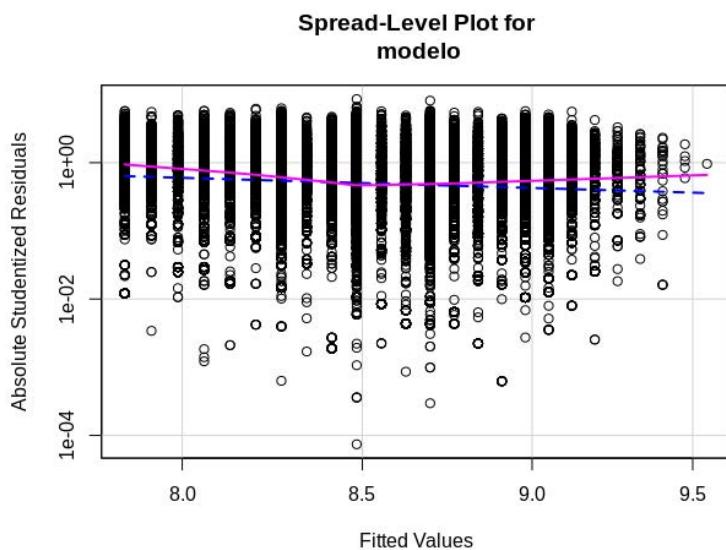
$H_a$ : Los errores no son homocedásticos

Nuestro modelo tiene un problema de heterocedasticidad puesto que hay evidencia estadística para rechazar la homocedasticidad de los errores al tener un valor- $p$  muy pequeño que nos hace rechazar la  $H_0$ . La heterocedasticidad crea un problema, ya que subestima los errores estándar con los que calculamos las pruebas de hipótesis para la estimación de los coeficientes de la recta.

Esto también se puede evaluar de manera gráfica con la gráfica vi-4. En el eje horizontal se establecen los valores predichos  $\hat{y}$ , mientras que en el eje vertical van los residuos. En general, quisiéramos que no hubiera ningún tipo de relación y observar algo más bien parecido a una nube de puntos:

```
# plot studentized residuals vs. fitted values
spreadLevelPlot(modelo)
```

**Gráfica vi-4. Diagrama de nivel de dispersión de los valores ajustados contra los residuos absolutos t-Student**



FUENTE: elaboración propia con base en la ENOE.

```
## 
## Suggested power transformation: 3.915714
```

En el caso del supuesto 4, asunción de normalidad, primero vamos a observar si existen datos atípicos que afecten la forma de nuestra distribución. Esto lo haremos con un test de detección de valores atípicos, el cual los lista según el número de su fila:

```
# Assessing Outliers
outlierTest(modelo) # Bonferroni p-value for most extreme obs

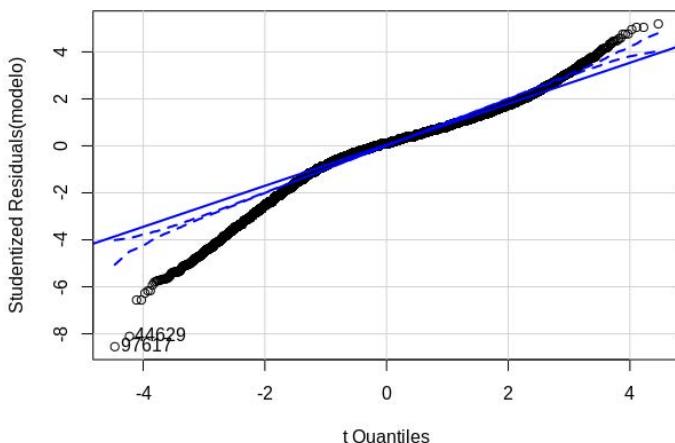
##          rstudent unadjusted p-value Bonferroni p
## 97617    -8.549673  1.2479e-17  1.5678e-12
## 44629    -8.109281  5.1373e-16  6.4540e-11
## 112654   -6.563986  5.2591e-11  6.6070e-06
## 114113   -6.563986  5.2591e-11  6.6070e-06
## 113598   -6.270182  3.6179e-10  4.5452e-05
## 108616   -6.172255  6.7528e-10  8.4834e-05
## 113650   -6.166008  7.0248e-10  8.8251e-05
## 109621   -5.931066  3.0176e-09  3.7910e-04
## 70394    -5.803493  6.5103e-09  8.1788e-04
## 116134   -5.776783  7.6321e-09  9.5882e-04
```

```
out<-outlierTest(modelo) # guardamos en objeto
```

Una manera de saber si una distribución es normal se puede usar un método gráfico como diagnóstico de diferencias entre la distribución del logaritmo de los ingresos válidos y una distribución normal. Ello se va comparando cuantil por cuantil, tal como se observa en la Gráfica VI-5.

```
qqPlot(modelo) #qq plot for studentized resid
```

**Gráfica VI-5. Diagrama Q-Q de distribución del primer modelo**



FUENTE: elaboración propia con base en la ENOE.

```
## [1] 44629 97617
qqPlot<-qqPlot(modelo) #guardamos en objeto
```

Si nuestra distribución fuera normal, tendríamos que ver cómo las dos rectas se superponen. Pero vemos que —sobre todo en los extremos— esto no es así. Existen varias diferencias. Del mismo modo, la gráfica arroja dos observaciones (líneas 44 629 y 97 617) que son las que más se alejan y que podrían ser las más problemáticas por su cualidad de ser tan atípicas.

Una decisión metodológica a considerar es eliminar los valores atípicos para nuestro ajuste; algo que analizaremos con detalle. Primero, volvemos a correr nuestro modelo con una base que elimine estas observaciones. Para ello, nos servirán los objetos que guardamos como resultado.

Creamos un objeto que liste los casos atípicos, como sigue:

```
names(out$bonf.p)
## [1] "97617" "44629" "112654" "114113" "113598" "108616" "113650"
## [8] "109621" "70394" "116134"
as.integer(names(out$bonf.p))
## [1] 97617 44629 112654 114113 113598 108616 113650 109621 70394 116134
outliers<-rbind(as.integer(names(out$bonf.p)), qqPlot) # Lista los casos
```

Vamos a eliminar estos casos que son extremos —no hay que olvidar que en una situación real esta decisión tiene implicaciones de interpretación y debe ser justificada metodológicamente, señalada como una limitante en un estudio—. Hemos dejado un grupo de personas que representaban a tanto más en nuestra población, simplemente porque no se ajustan a nuestro modelo. Es decir, nuestro modelo no los explica.

Tenemos el nombre de las filas que nos dan problemas en el objeto *outliers* y vamos a deseleccionar estos casos y crear un nuevo objeto para correr nuevamente nuestro modelo:

```
dmodelo$rownames<-rownames(dmodelo)
dmodelo2<-dmodelo[-outliers,]
```

Corremos un nuevo modelo y guardamos sus resultados en un objeto llamado “modelo0”.

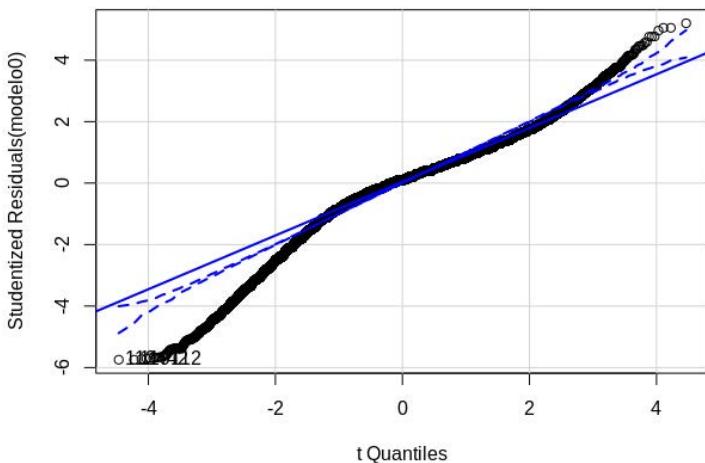
```
modelo0<-lm(log_ingroup ~anios_esc, data=dmodelo2, na.action=na.exclude)
summary(modelo0)

##
## Call:
## lm(formula = log_ingroup ~ anios_esc, data = dmodelo2, na.action = na.exclude)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -4.1478 -0.3303  0.0751  0.4307  3.7532 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 7.8475899  0.0053308   1472   <2e-16 ***
## anios_esc   0.0707590  0.0004847    146   <2e-16 ***  
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 0.722 on 125617 degrees of freedom
## Multiple R-squared:  0.145, Adjusted R-squared:  0.145 
## F-statistic: 2.131e+04 on 1 and 125617 DF, p-value: < 2.2e-16
```

Si vemos, nuestro ajuste no cambió demasiado. Revisemos si eliminar estos casos resolvió nuestro problema de normalidad:

```
qqPlot(modelo0)
```

**Gráfica vi-6. Diagrama Q-Q de distribución del modelos sin valor atípicos**



FUENTE: elaboración propia con base en la ENOE.

```
## 113442 119112
## 113436 119102

outlierTest(modelo0)

##          rstudent unadjusted p-value Bonferroni p
## 113442 -5.746044   9.1566e-09   0.0011502
## 119112 -5.746044   9.1566e-09   0.0011502
## 114187 -5.726953   1.0248e-08   0.0012874
## 73043 -5.707331   1.1502e-08   0.0014449
## 113448 -5.693561   1.2469e-08   0.0015664
## 84221 -5.683002   1.3264e-08   0.0016662
## 72897 -5.680836   1.3433e-08   0.0016875
## 52587 -5.667947   1.4483e-08   0.0018194
## 122882 -5.667947   1.4483e-08   0.0018194
## 35533 -5.648787   1.6193e-08   0.0020342
```

Revisamos y seguimos teniendo valores atípicos. ¿Cuándo podría parar este proceso? ¡Este puede ser un proceso infinito! Si quitamos lo anormal, esto mueve nuestros rangos y, al quitar un *outlier*, otra observación que antes no era *outlier* en el ajuste se puede convertir en dato atípico. Por ello, quien lee este texto debe considerar cuidadosamente qué hacer ante la presencia de valores atípicos.

## C. Regresión lineal múltiple

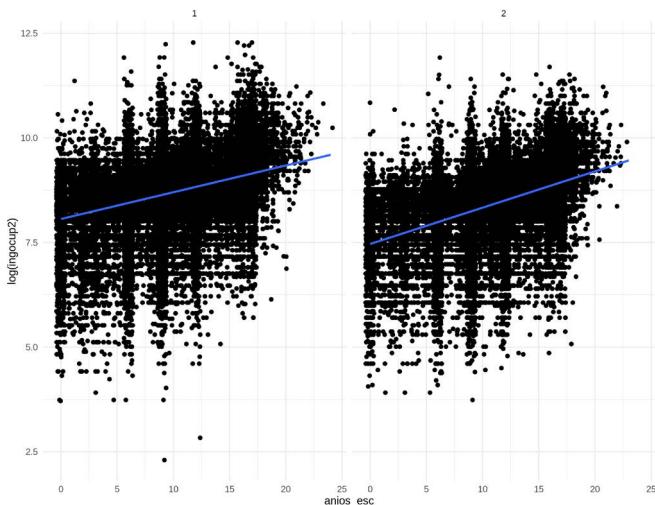
En realidad, es difícil que una sola variable explique por completo a otra. Como vimos en nuestro primer modelo, los años de escolaridad pueden predecir la varianza del logaritmo de los ingresos en un 14.5%. En ciencias sociales, además, las relaciones son complejas y múltiples. En esta sección iremos complicando el modelo al introducir más variables.

### a. Agregando una variable categórica

Con lo estudiado hasta el momento sobre el comportamiento de los mercados de trabajo mexicanos, hemos establecido que la dinámica de los ingresos es diferente entre hombres y mujeres. Esto lo podemos visualizar en una gráfica donde comparemos las rectas ajustadas para los grupos:

```
gg <- dmodelo %>%
  ggplot(aes(anios_esc, log(ingocup2))) # el Lienzo
  gg + geom_jitter() + #la geometría jitter
  geom_smooth(method="lm") + # geometría que ajusta una Línea recta (Lm=Linear model)
  facet_wrap(~as_label(sexo)) + # separa por categoría de sexo
  theme_minimal() # cambia apariencia
```

**Gráfica vi-7. Diagrama de dispersión de la relación entre los ingresos y los años de escolaridad. Hombres y Mujeres. México, trimestre III, 2019. Personas con remuneraciones válidas y mayores que cero**



FUENTE: elaboración propia con base en la ENOE.

Las líneas de cada panel son diferentes. Si bien parece que no lo son tanto en la inclinación, sí en el intercepto; es decir, el valor de las condiciones iniciales cuando la escolaridad es igual a cero. Cuando nosotros tenemos una variable categórica, como la condición de sexo, podemos modificar el intercepto en nuestro modelo. El modelo se escribe de la siguiente forma:

$$y = \beta_0 + \beta_1 x I + \delta_o \text{sexo} + \epsilon$$

Donde  $x$  son los años de escolaridad. Sexo debe ser codificada como una variable dicotómica o dummy, que sólo toma valores de 1 y 0. Si asumimos 1 cuando se es mujer y 0 cuando se es hombre, podemos ver que esta estimación sólo cambió en el intercepto.

De tal cuenta que la predicción para las mujeres resulta como:

$$y = \beta_0 + \beta_1 x + \delta_o * I + \epsilon$$

$$y = (\beta_0 + \delta_o) + \beta_1 x + \epsilon$$

El coeficiente calculado se puede establecer como un cambio en el intercepto, o un cambio entre ser mujer con respecto a ser hombre, manteniendo la educación constante. Este último elemento es indispensable. En una visión múltiple, las rectas sólo pueden ser descritas mientras se mantiene constante la otra u otras variables.

Vamos a introducir la variable al modelo. Seguimos haciendo el ejercicio, a pesar de que ya observamos en nuestro diagnóstico que el modelo no cumple con los supuestos. Lo haremos para fines ilustrativos. Introduzcamos el cambio:

```
dmodelo<-dmodelo %>% mutate(sex=as_label(sex)) #para guardar etiquetas
model01<-lm(log_ingocup ~ anios_esc + sex, data=dmodelo, na.action=na.exclude)
summary(model01)

##
## Call:
## lm(formula = log_ingocup ~ anios_esc + sex, data = dmodelo, na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.3257 -0.3143  0.0745  0.4312  3.8733
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.9667912  0.0053452 1490.47 <2e-16 ***
## anios_esc   0.0735018  0.0004716 155.85 <2e-16 ***
## sexMujer    -0.3627518  0.0040348 -89.91 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.701 on 125626 degrees of freedom
## Multiple R-squared:  0.1964, Adjusted R-squared:  0.1964
## F-statistic: 1.536e+04 on 2 and 125626 DF,  p-value: < 2.2e-16
```

Este modelo tiene coeficientes que deben leerse “condicionados”. Es decir, en este caso tenemos que cuando analizamos el coeficiente asociado a “anios\_esc”, se mantiene constante el valor de sexo y viceversa. Manteniendo la escolaridad constante, el cambio por ser una mujer trabajadora es de  $-36.28\%$ , en los ingresos laborales, con respecto a un hombre.

¿Cómo saber si ha mejorado nuestro modelo con la introducción de una nueva variable? Podemos comparar el ajuste con la ANOVA; es decir, con una prueba  $F$ , la cual compara las varianzas explicadas por ambos modelos.

```
pruebaf0<-anova(modelo, modelo1)
pruebaf0

## Analysis of Variance Table
##
## Model 1: log_ingocup ~ anios_esc
## Model 2: log_ingocup ~ anios_esc + sex
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1 125627 65707
## 2 125626 61735  1     3972.1 8082.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los resultados reportan un grado de libertad de 1 (lo que indica que el modelo más complejo tiene un parámetro adicional) y un valor-*p* muy pequeño (<.001). Esto significa que agregar el sexo al modelo lleva a un ajuste significativamente mejor sobre el modelo original.

## b. Un modelo más complejo

Podemos seguir añadiendo variables “sumando” en la función.

```
modelo2<-lm(log_ingocup ~ anios_esc + sex + eda, data=dmodelo, na.action=na.exclude)
summary(modelo2)

##
## Call:
## lm(formula = log_ingocup ~ anios_esc + sex + eda, data = dmodelo,
##      na.action = na.exclude)
##
## Residuals:
##       Min     1Q     Median      3Q     Max 
## -6.3100 -0.3135  0.0827  0.4204  3.8296 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 7.7802748  0.0088311 881.00 <2e-16 ***
## anios_esc   0.0770800  0.0004893 157.53 <2e-16 ***
## sexMujer   -0.3655882  0.0040251 -90.83 <2e-16 ***
## eda         0.0038484  0.0001453  26.49 <2e-16 ***  
## ---
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6991 on 125625 degrees of freedom
## Multiple R-squared:  0.2009, Adjusted R-squared:  0.2009 
## F-statistic: 1.053e+04 on 3 and 125625 DF, p-value: < 2.2e-16
```

Y vemos si introducir esta variable afectó al ajuste global del modelo.

```
pruebaf1<-anova(modelo1, modelo2)
pruebaf1

## Analysis of Variance Table
##
## Model 1: log_ingocup ~ anios_esc + sex
## Model 2: log_ingocup ~ anios_esc + sex + eda
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1 125626 61735
## 2 125625 61392  1     342.92 701.71 < 2.2e-16 ***
## ---
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ya que tenemos más variables, podemos revisar dos supuestos más.

Además de los supuestos de la regresión simple con más variables en nuestro modelo, debemos revisar si nuestro supuesto de linealidad en los parámetros se mantiene para todos. Además, revisamos si se distribuyen de manera normal, pero de forma multivariada.

Uno de los supuestos que nos da más problemas es la multicolinealidad, debido a que las variables explicativas están muy relacionadas entre sí. Para verificar esta situación, utilizamos el comando “vif()”, del paquete car, que ejecuta la prueba de factor influyente de la varianza (VIF por sus siglas en inglés). La lógica es que la multicolinealidad tendrá efectos en nuestro  $R^2$ , inflándolo. Ahí observamos de cuál o cuáles variables proviene este problema relacionado con la multicolinealidad.

Si el valor es mayor a 5, tenemos un problema muy grave de multicolinealidad, que podría estar afectando demasiado nuestro  $R^2$ .

```
vif(modelo2)
## anios_esc      sex      eda
##  1.086914  1.004794  1.082563
```

Un indicador de este problema es tener  $R^2$  muy altos y pocos coeficientes que sean estadísticamente significativos a niveles de significancia tradicionales, o bien, con intervalos de confianza muy amplios.

## D. Presentación de modelos

Para presentar los modelos estimados y que se puedan comparar fácilmente, usualmente se describen con tablas; donde hay espacio tanto para estimaciones de los coeficientes de cada variable como para los estadísticos de ajustes globales.

El paquete *stargazer* nos provee de un comando sencillo que, además, puede servir para comparar modelos a los que hemos ido sumando variables y podemos revisar su evolución.

```

stargazer(modelo, modelo1, modelo2, type = 'text', header=F)

##
## =====
##                               Dependent variable:
##                               log_ingocup
## -----
##             (1)                   (2)                   (3)
## -----
## anios_esc      0.071***      0.074***      0.077***  

##                 (0.0005)      (0.0005)      (0.0005)  

##  

## sexMujer      -0.363***      -0.366***  

##                 (0.004)       (0.004)  

##  

## eda           0.004***  

##                 (0.0001)  

##  

## Constant      7.847***      7.967***      7.780***  

##                 (0.005)       (0.005)       (0.009)  

##  

## Observations   125,629        125,629        125,629  

## R2            0.145          0.196          0.201  

## Adjusted R2    0.145          0.196          0.201  

## Residual Std. Error 0.723 (df = 125627) 0.701 (df = 125626) 0.699 (df = 125625)  

## F Statistic   21,259.970*** (df = 1; 125627) 15,355.270*** (df = 2; 125626) 10,527.850*** (df = 3; 125625)  

## =====
## Note: *p<0.1; **p<0.05; ***p<0.01

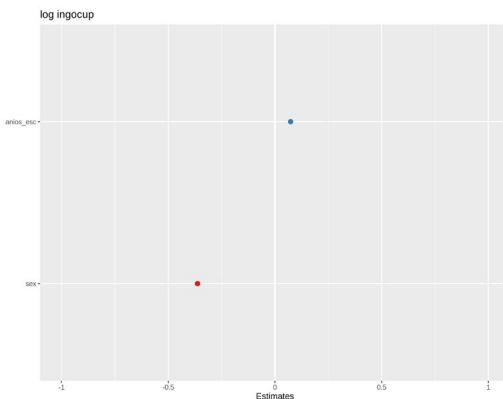
```

Esta manera de presentar los datos permite comparar fácilmente los modelos. Es muy importante establecer que la comparación entre los ajustes de los modelos implica que deben tener la misma cantidad de observaciones. En este cuadro podemos ver claramente cómo mejoró el ajuste. Cuando comparamos el  $R^2$  ajustado, vemos que pasamos de 14.5% de varianza explicada al 20.1%. Del mismo modo, el error estándar ha disminuido. Ello muestra que la introducción de la variable de sexo y edad mejora los ajustes de nuestro modelo.

También podemos representar los coeficientes de manera gráfica. Normalmente, lo que buscamos es comparar contra el cero. El paquete *sjPlot*, con el comando “*plot\_model()*”, nos permite graficar un modelo específico (véase Gráfica VI-8, en página siguiente).

```
plot_model(modelo1)
```

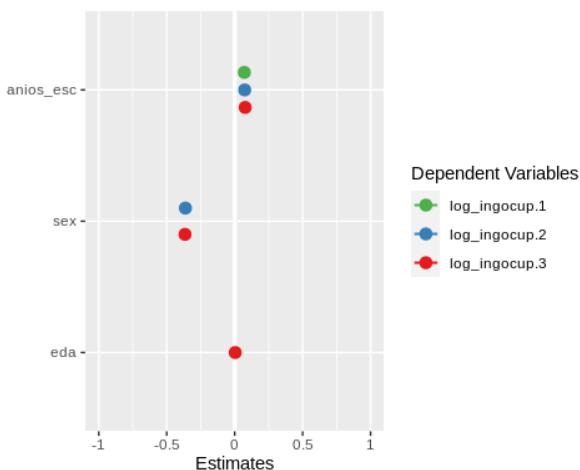
La Gráfica VI-8 muestra claramente cómo los años de escolaridad están a la derecha del cero, marcando un efecto positivo y distinto de cero. Mientras que la variable categórica “mujer” está del lado izquierdo, marcando un efecto negativo.

**Gráfica vi-8. Coeficientes del modelo 1**

FUENTE: elaboración propia con base en la ENOE.

Si queremos graficar varios modelos en una sola gráfica, utilizamos el comando “plot\_models()” y dentro de los argumentos ponemos los objetos con los resultados del modelo:

```
plot_models(modelo, modelo1, modelo2)
```

**Gráfica vi-9. Coeficientes del modelo 1, modelo 2 y modelo 3**

FUENTE: elaboración propia con base en la ENOE.

Estas gráficas son compatibles con *ggplot2*, por lo que muchos comandos para editarlos se pueden realizar de manera similar que en *ggplot* y también se pueden guardar en objetos para su posterior edición.

## E. Estandarizando las unidades de medida las variables

La comparación de los resultados de los coeficientes es difícil porque el efecto está en términos de las unidades de medida. En nuestro modelo no sería tan comparable el efecto que tenemos de los años de escolaridad (que se mide en años en un sistema escolar) con respecto a la edad (que se mide en años vividos); o bien, ¿cómo comparar el hecho de ser mujer o no con las otras dos variables? Por ello, las gráficas VI-8 y VI-9 no permiten hacer comparaciones del tamaño de los efectos, sino sólo de sentido y de su relación respecto a 0.

Para poder comparar efectos, se necesita transformar las variables de origen e introducirlas al modelo en una nueva escala. Decimos estandarizar porque el proceso proviene de pasar cada una de las mediciones a su puntaje “Z”, es decir, restando la media y dividiendo entre la desviación estándar. Lo que observaremos en los coeficientes estimados lo podemos entender como los cambios en la variable dependiente, dado un aumento en una desviación estándar de la variable independiente.

En el caso de *R*, podemos usar un paquete que realiza la transformación directamente. A los coeficientes calculados se les conoce como “beta”.

Simplemente aplicamos el comando a nuestros modelos ya calculados.

```
lm.beta(modelo2)

##
## Call:
## lm(formula = log_ingocup ~ anios_esc + sex + eda, data = dmodelo,
##     na.action = na.exclude)
##
## Standardized Coefficients:
## (Intercept)    anios_esc    sexMujer        eda
## 0.00000000  0.41420079 -0.22962539  0.06951304
```

La comparación será mucho más clara y podemos ver qué variable tiene mayor efecto en nuestra variable dependiente.

```

modelo_beta<-lm.beta(modelo2)
modelo_beta

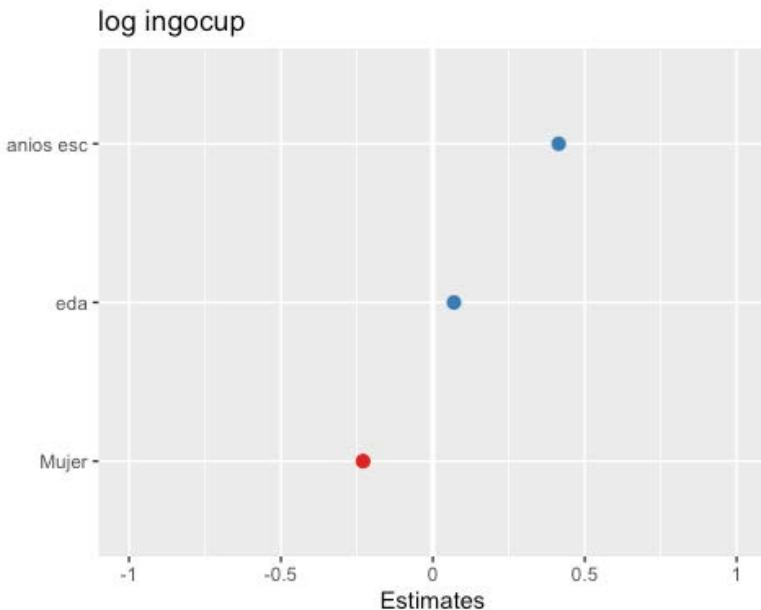
##
## Call:
## lm(formula = log_ingocup ~ anios_esc + sex + eda, data = dmodelo,
##     na.action = na.exclude)
##
## Standardized Coefficients:
## (Intercept)  anios_esc      sexMujer       eda
##  0.00000000  0.41420079 -0.22962539  0.06951304

```

Para graficarlos, podemos usar de nuevo el comando “plot\_model()” con una opción “std”.

```
plot_model(modelo2, type="std")
```

### Gráfica vi-10. Coeficientes del modelo 1. Coeficientes estandarizados



FUENTE: elaboración propia con base en la ENOE.

Si comparamos la Gráfica vi-9 con la vi-10, notamos que la edad no parece tener un efecto tan pequeño como cuando teníamos que interpretar en términos de años cumplidos. Además, podemos decir que el efecto estandarizado indica que un cambio en la desviación estándar de los años de escolaridad aumenta casi 40% los niveles salariales, manteniendo lo demás constante. Queda claro que los años de escolaridad tienen un efecto muy importante en la predicción de las remuneraciones.

## F. No cumplo los supuestos ¿Y ahora qué?

Iniciamos este capítulo con una introducción a los modelos de regresión. Con el avance programático, los modelos estadísticos se han ido refinando y las operaciones complejas se pueden hacer con más facilidad. De ahí que existan modificaciones al modelo de regresión que nos permiten lidiar con problemas que no cumplen con los supuestos. A continuación veremos lo que se puede hacer en el caso de la heterocedasticidad (un problema muy común en análisis con información de corte transversal) y qué hacer cuando se presentan muchos valores atípicos.

### a. Heterocedasticidad

El problema de la heterocedasticidad es que los errores estándar de los coeficientes se subestiman, por lo que si estos están en el cociente del estadístico de prueba t implica que nuestras pruebas podrían arrojar valores significativos cuando no lo son.

Una forma muy sencilla es pedir los errores robustos; para lo cual habrá que importar una función de la siguiente liga: <<https://economictheoryblog.com/2016/08/08/robust-standard-errors-in-r/>>.

```
library(RCurl)
url_robust <- "https://raw.githubusercontent.com/IsidoreBeautrelet/economictheoryblog/
master/robust_summary.R"
eval(parse(text = getURL(url_robust, ssl.verifypeer = FALSE)),
     envir=.GlobalEnv)
```

Una vez que hayan corrido esas líneas, se ejecuta un “summary()” al modelo al que queramos ver con errores robustos, con el estimador “sandwich” de Huber y White:

```
summary(modelo2, robust=T)
##
## Call:
## lm(formula = log_ingroup ~ anios_esc + sex + eda, data = dmodelo,
##     na.action = na.exclude)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -6.3100 -0.3135  0.0827  0.4204  3.8296 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 7.7802748  0.0092277 843.14   <2e-16 ***
## anios_esc   0.0770800  0.0005308 145.21   <2e-16 ***
## sexMujer   -0.3655882  0.0041025 -89.11   <2e-16 ***  
## eda        0.0038484  0.0001621  23.74   <2e-16 ***  
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6991 on 125625 degrees of freedom
## Multiple R-squared:  0.2009, Adjusted R-squared:  0.2009 
## F-statistic:  9018 on 3 and 125625 DF,  p-value: < 2.2e-16
```

## b. Regresión robusta a valores atípicos

Cuando tenemos muchos valores atípicos, podemos hacer una regresión donde se consideren pesos diferenciados a las observaciones, de tal forma que estos valores atípicos no influyan demasiado en los resultados. Esto se hace con diversos algoritmos. En este caso, presentamos un ejemplo desarrollado con el paquete *robustbase* donde:

[...] esta función calcula un estimador de regresión de tipo MM como se describe en Yohai (1987) y Koller y Stahel (2011). De forma predeterminada, utiliza una función de puntuación de redistribución de dos cuadrados y devuelve un estimador muy robusto y muy eficiente (con un punto de ruptura del 50% y un 95% de asintóticas) (Maechler *et al.*, 2020).

Tenemos que volver a estimar el modelo. Los comandos y sus resultados se muestran a continuación:

```
modelo2rob<-lmrob(log_ingroup ~ anios_esc + sex + eda, data=dmodelo, na.action=na.exclude)
summary(modelo2rob)

##
## Call:
## lmrob(formula = log_ingroup ~ anios_esc + sex + eda, data = dmodelo, na.action = na.exclude)
##   \--> method = "MM"
## Residuals:
##   Min     1Q Median     3Q    Max 
## -6.373750 -0.383711  0.009592  0.342171  3.669828 
##
```

```

## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.8775422  0.0000492 978.68 <2e-16 ***
## anios_esc   0.0683610  0.0004905 139.36 <2e-16 ***
## sexMujer    -0.3195610  0.0035666 -89.60 <2e-16 ***
## eda         0.0050985  0.0001392  36.62 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Robust residual standard error: 0.5285
## Multiple R-squared:  0.2273, Adjusted R-squared:  0.2273
## Convergence in 11 IRLWLS iterations
##
## 10821 weights are ~ 1. The remaining 113749 ones are summarized as
##   Min. 1st Qu. Median Mean 3rd Qu. Max.
## 0.0000016 0.8492000 0.9498000 0.8718000 0.9859000 0.9990000
## Algorithmic parameters:
##      tuning.chi          bb      tuning.psi      refine.tol
##      1.548e+00      5.000e-01      4.685e+00      1.000e-07
##      rel.tol        scale.tol      solve.tol      eps.outlier
##      1.000e-07      1.000e-10      1.000e-07      7.960e-07
##      eps.x warn.limit.reject warn.limit.meanrw
##      1.783e-10      5.000e-01      5.000e-01
##      nResample     max.it      best.r.s      k.fast.s      k.max
##      500           50            2             1            200
##      maxit.scale   trace.lev      mts      compute.rd fast.s.large.n
##      200            0           1000            0            2000
##      psi          subsampling      cov
##      "bisquare"    "nonsingular" ".vcov.avar1"
##      compute.outlier.stats
##      "SM"
##      seed : int(0)

```

Se debe recordar que no es lo mismo la regresión robusta que los errores estándar robustos. La regresión robusta se llama así por ser robusta a los *outliers*.

Comparemos en una tabla esta estimación de regresión robusta contra una de MCO:

```

stargazer(modelo2, modelo2rob, type = 'text', header=FALSE)

##
## =====
##                               Dependent variable:
##                               -----
##                               log_ingroup
##                               OLS      MM-type
##                               linear
##                               (1)      (2)
## -----
## anios_esc                      0.077***      0.068***  

##                               (0.0005)     (0.0005)  

##  

## sexMujer                     -0.366***      -0.320***  

##                               (0.004)       (0.004)  

##  

## eda                          0.004***      0.005***  

##                               (0.0001)     (0.0001)  

##  

## Constant                      7.780***      7.878***  

##                               (0.009)      (0.008)

```

```

## -----
## Observations           125,629          125,629
## R2                   0.201           0.227
## Adjusted R2           0.201           0.227
## Residual Std. Error (df = 125625)   0.699           0.529
## F Statistic          10,527.850*** (df = 3; 125625)
## =====
## Note:                *p<0.1; **p<0.05; ***p<0.01

```

Vemos que la regresión robusta mejora el ajuste, puesto que las observaciones atípicas ya no ejercen tanta influencia. También observamos cambios en nuestros coeficientes pero sin cambios drásticos de sentido. Para saber más de la regresión robusta, también puede leerse lo que se ha desarrollado para otros paquetes (Yaffee, 2002).

Además de estas opciones, cada vez se están usando más métodos no paramétricos, como la regresión mediana; ello se puede llevar a cabo con el paquete *quantreg* (Koenker, 2020). También, con este tipo de modelos, se pueden modelar otros cuantiles distintos a la mediana. Esto puede ser muy interesante, sobre todo cuando se estudian mercados de trabajo segmentados (Márquez, Prieto & Escoto, 2020).

En los ejemplos de este capítulo quitamos los ingresos iguales a cero. Esto resuelve un problema estadístico, pero nos supone una decisión metodológica difícil de atender: ¿Cómo estudiamos a los que no tienen ingresos? ¿Cómo los incluimos? ¿Cómo hacemos explícita esta situación? Existen algunos tipos de métodos que permiten modelar esto, como la corrección de Heckman, y otras extensiones para corregir por selectividad de la población de la muestra; así como modelos de tipo Tobit que suponen un tipo de censura en los datos. Se puede revisar el paquete *sampleSelection* para revisar estos métodos (Henningsen, Toomet & Petersen, 2019).

Del mismo modo, hay extensiones de los modelos que nos permiten no asumir linealidad en los parámetros. Variables como la edad y los años de escolaridad, normalmente no tienen un comportamiento lineal y muchas veces se modelan con efectos cuadráticos que nos permiten hablar de si hay un proceso que se acelera o se desacelera.

Estos son algunos ejemplos de otros métodos a utilizar que le darán más robustez a los análisis. A pesar de algunas limitantes que se nos presentaron, en este capítulo seguimos ahondando en el estudio de las disparidades en el mercado de trabajo. Las mujeres, aun

si tienen la misma edad y los mismos años de escolaridad que los hombres, mantienen remuneraciones mucho más bajas.

Estas brechas ya han sido sumamente analizadas y también se ha estudiado cómo se descomponen a través de la metodología Oaxaca-Blinder (Ospino, Roldán & Barraza, 2010), que se puede llevar a cabo con el paquete *oaxaca* (Hlavac, 2018a).

Los modelos que revisamos son sencillos e incluyen pocas variables. Sin duda, la potencia de la ENOE despierta muchas preguntas de investigación, tal y como lo vimos en el capítulo 1. Si bien nos da estimaciones transversales, tiene un uso longitudinal que ha sido menos explorado. Eso revisaremos en el capítulo final de este libro.

## **VII. Aplicaciones longitudinales con la ENOE. El caso del análisis de secuencias**

### **Introducción**

Como señalamos en el capítulo II, la ENOE tiene un diseño de panel rotativo. Esto quiere decir que se siguen a personas a quienes se les administra la encuesta en varios momentos en el tiempo. En específico, en la ENOE se siguen a las personas en cinco ocasiones. Posteriormente, se dejan de entrevistar. Esto significa que, en la sexta ocasión, su puesto es reemplazado por alguien más.

Esto permite establecer análisis donde podemos observar las respuestas y estrategias de los trabajadores; o bien, estudiar una dimensión que se ha estudiado desde la perspectiva de la precariedad laboral y las carencias laborales: la inestabilidad de la inserción (García, 2011). Algunos trabajos de este tipo ya se venían haciendo con la predecesora de la ENOE, la eneu, y varios responden a la manera en que cambia la inserción laboral a partir de las crisis que ha vivido el país, en específico con el análisis de tasas de transición y construcción de trayectorias (Flores, Zamora & Contreras, 2013; Freije, López & Rodríguez, 2011; Ochoa, 2016; Pacheco & Parker, 2001; Partida, 2011; Salas, 2003). Del mismo modo, esta mirada longitudinal ha permitido estudiar trayectorias particulares de las mujeres (Escoto, 2020; García, 2014).

Para aproximarnos al uso del panel y algunas de sus aplicaciones, el capítulo se ha estructurado en cuatro secciones. En la primera se reseña el diseño longitudinal de la ENOE y se establece el proceso de cómo se construye el panel para crear una base de datos longitudinal: el seguimiento de los individuos y su identificación. Se ejemplifica con el código replicable en R, utilizando a los individuos que se entrevistaron desde el cuarto trimestre de 2018 hasta el cuarto

trimestre de 2019. En una segunda sección, se discuten los formatos de trabajo de una base de datos longitudinal. Posteriormente, en las dos últimas secciones, se presentan dos ejemplos longitudinales: el análisis de secuencias y las tasas de transición.

## A. Construcción del panel de la ENOE

Para comenzar, vamos a retomar el fusionado que hicimos en el capítulo II y replicarlo en cuatro ocasiones más para así construir el panel. El código y el ambiente están disponibles en <<https://tinyurl.com/enoe-panel>>. Repliquemos ahora las siguientes operaciones:

```
if (!require("pacman")) install.packages("pacman") # instala pacman si se requiere
## Loading required package: pacman
#cargan Los paquetes necesarios para la práctica de este capítulo
pacman::p_load(tidyverse,haven,sjlabelled, janitor, extdplyr)

load("panel.RData") # se descarga este archivo en la carpeta de trabajo que has establecido
ls() # Listamos Los objetos de este ambiente: tenemos cinco bases ya fusionadas

## [1] "completat119" "completat219" "completat319" "completat418"
## [5] "completat419"
```

Tenemos cinco bases fusionadas de la ENOE; con ellas reconstruiremos el panel desde el trimestre IV de 2018. Esto significa que los individuos a quienes se dio seguimiento y se entrevistaron en ese trimestre tuvieron su quinta entrevista en el trimestre IV de 2019.

Si tabulamos el número de individuos por entrevista, veremos que más o menos los individuos se dividen equitativamente según la entrevista y que, en cada edición, hay individuos que han realizado diferentes cantidades de entrevistas:

```
completat418 %>%
  mutate(n_ent=as_label(n_ent)) %>%
  tabyl(n_ent)

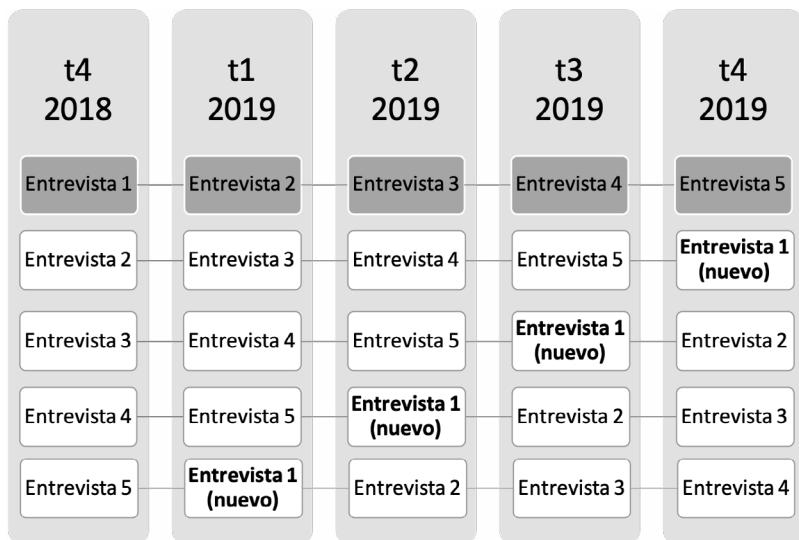
##          n_ent      n   percent
##  Primera entrevista 77134 0.2011789
##  Segunda entrevista 77595 0.2023813
##  Tercera entrevista 76701 0.2000496
##  Cuarta entrevista 75985 0.1981821
##  Quinta entrevista 75995 0.1982082

completat419 %>%
  mutate(n_ent=as_label(n_ent)) %>%
  tabyl(n_ent)
```

```
##           n_ent      n   percent
## Primera entrevista 78846 0.1994460
## Segunda entrevista 79294 0.2005793
## Tercera entrevista 79837 0.2019528
## Cuarta entrevista 79073 0.2000202
## Quinta entrevista 78275 0.1980016
```

Al llegar a la entrevista número cinco, la vivienda se sustituye; de ahí su carácter rotativo. En la figura que sigue observamos cuál es la lógica de este panel y qué construiremos con este código. El panel corresponde a las figuras en gris oscuro.

**Figura VII-1. Estructura del panel de la ENOE**



En cada una de las bases podemos encontrar la variable “n\_ent”, que refiere al número de entrevista. Al añadir el número de entrevista al identificador de la persona, tendremos el identificador único que combina el número de entrevista con cada uno de los individuos.

```
idpersona<-c("cd_a", "ent", "con", "v_sel","n_hog", "h_mud", "n_ren")
idpanel<-c("cd_a", "ent", "con", "v_sel","n_hog", "h_mud", "n_ren", "n_ent")
```

Podemos revisar que sólo existe una observación por persona en cada una de las bases:

```
completat119 %>%
  get_dupes(cd_a, ent, con, v_sel,n_hog, h_mud, n_ren) %>% #genera número de duplicados
  tabyl(dupe_count) #Nos dice que no hay duplicados

## No duplicate combinations found of: cd_a, ent, con, v_sel, n_hog, h_mud, n_ren

## [1] dupe_count n      percent
## <0 rows> (or 0-length row.names)
```

Como tenemos la base completa, esto puede resultar bastante pesado para las computadoras con capacidades de memoria limitadas. Para hacer este ejercicio, vamos a utilizar la información que provenga únicamente del cuestionario sociodemográfico. Sin embargo, hemos incluido las bases de los cinco paneles para que quien consulte este libro pueda escoger las variables que le parezcan relevantes.

```
ejt418 <- completat418[, 1:120]
ejt119 <- completat119[, 1:120]
ejt219 <- completat219[, 1:120]
ejt319 <- completat319[, 1:120]
ejt419 <- completat419[, 1:120]

rm(completat418, completat119, completat219, completat319, completat419 )
```

Es útil crear un identificador para la temporalidad de cada variable; así, a la hora de realizar el panel, tendremos identificado no sólo el número de entrevista, sino también cuándo fue administrada. Seguiremos a nuestros individuos, por lo que en cada trimestre nos quedaremos únicamente con el número de entrevista que nos interesa para tener nuestro panel completo.

```
ejt418 <- ejt418 %>%
  filter(n_ent==1) %>%
  mutate(trim="t418")

ejt119 <- ejt119 %>%
  filter(n_ent==2) %>%
  mutate(trim="t119")

ejt219 <- ejt219 %>%
  filter(n_ent==3) %>%
  mutate(trim="t219")

ejt319 <- ejt319 %>%
  filter(n_ent==4) %>%
  mutate(trim="t319")

ejt419 <- ejt419 %>%
  filter(n_ent==5) %>%
  mutate(trim="t419")
```

Esta variable la incluiremos en nuestro identificador:

```
idpanel<-c("cd_a", "ent", "con", "v_sel","n_hog", "h_mud", "n_ren", "n_ent", "trim")
```

Usamos el comando “merge()” pero incorporando todos los nombres de nuestra base de datos. Al escribir la opción “all=T” unirá la información de todas las bases.

```
panel<-ejt418 %>%
  merge(ejt219, by=names(ejt418), all=T) %>%
  merge(ejt319, by=names(ejt418), all=T) %>%
  merge(ejt419, by=names(ejt418), all=T) %>%
  merge(ejt119, by=names(ejt418), all=T) # dejar éste de último
```

Se debe recordar que el primer trimestre tiene la particularidad de que se aplica un cuestionario diferente y, por lo tanto, tiene variables distintas. Si nuestras variables fueran idénticas, podríamos crear un objeto con el comando “rbind()”. En este caso, nosotros tenemos variables iguales porque nos quedamos con las que aporta el cuestionario sociodemográfico, el cual es común.

No obstante, hay que tener cuidado al pegar la información del cuestionario del trimestre 1 y revisar qué variables diferentes aporta esta base de datos. Por ello utilizamos esta forma con “merge ()”. En el argumento “by=” se debe colocar el nombre de las variables comunes a las bases. Aconsejamos pegar primero todas las bases que no son del cuestionario ampliado y luego hacer el “merge ()” con el primer trimestre del año.

Para liberar memoria de nuestro ambiente, quitamos las bases individuales.

```
rm(ejt119, ejt219, ejt319, ejt418, ejt419)
```

Revisamos los identificadores.

```
panel %>%
  get_dups(cd_a, ent, con, v_sel,n_hog, h_mud, n_ren, n_ent, trim) %>%
# persona y entrevista
  tabyl(dupe_count) # no hay duplicados, ningún individuo está entrevistado más de una
ocasión en el mismo trimestre

## No duplicate combinations found of: cd_a, ent, con, v_sel, n_hog, h_mud, n_ren, n_en
t, trim

## [1] dupe_count n      percent
## <0 rows> (or 0-length row.names)
```

Si quitamos la variable “trim”, observaremos que todas las personas aparecen en más de una ocasión, pero no todas aparecen las cinco veces.

```
panel %>%
  get_dupes(cd_a, ent, con, v_sel,n_hog, h_mud, n_ren) %>%
  tabyl(dupe_count) # tabula Los duplicados

### dupe_count      n    percent
###              2 16560 0.04348351
###              3 21726 0.05704848
###              4 45308 0.11897047
###              5 297240 0.78049754
```

Para poder agregar esto como una variable, utilizamos el comando “add\_tally()” y así nos incluye como variable la cantidad de observaciones. La ENOE, como se estableció en el capítulo II, tiene un informante clave por hogar. Ello quiere decir que la información puede cambiar debido a que quien la reporta es diferente; o podría haber sucedido alguna sustitución excepcional. Sin embargo, existen atributos de una persona que no cambian a lo largo de cinco trimestres. Por eso podemos agregar “sexo” como una variable de control. Tampoco sería posible la modificación de la edad en más de dos años durante un trimestre. Incluiríremos estas dos variables para verificar que se trata de los mismos individuos.

```
#creación de variable de discrepancia de edad

panel<-panel %>%
  group_by(cd_a, ent, con, v_sel,n_hog, h_mud, n_ren) %>% # agrupa Las
  observaciones por persona
  mutate(edad_min=min(eda), edad_max=max(eda)) %>% #variables de edad
  máxima y mínima de cada individuo
  mutate(edad_diff=edad_max-edad_min) %>% # variable de diferencia
  ungroup() # para que no nos guarde la variable como un grupo

#conteo de observaciones por persona

panel<- panel %>%
  filter(edad_diff<2 & edad_diff>-2) %>% # filtra los casos que tienen mucha discrepancia
  de edad
  group_by(cd_a, ent, con, v_sel,n_hog, h_mud, n_ren, sex) %>%
  add_tally() %>% # agrega el resultado a nuestro objeto como una variable
  rename(total_n=n) %>% # le cambiamos el nombre de n
  ungroup() # para que no nos guarde la variable como un grupo

panel %>% tabyl(total_n) ## se pierden observaciones. Hay quienes sólo participaron una vez
```

```
##   total_n      n    percent
##     1 12799 0.03261307
##     2 16560 0.04219646
##     3 21726 0.05535992
##     4 45280 0.11537776
##     5 296085 0.75445280
```

Si sumamos el número de casos que no tienen las cinco observaciones, podemos calcular la atracción. Observamos que la atracción es alta, ya que perdemos casi 25% de los individuos.

```
panel %>% mutate(atencion=total_n==5) %>%
  tabyl(atencion)

## #> #> #> atricion      n    percent
## #> #> FALSE 96365 0.2455472
## #> #> TRUE 296085 0.7544528
```

Para poder trabajar con el panel, seleccionamos únicamente los que tienen cinco observaciones, es decir, el panel completo. En esta nueva base de datos, los individuos tienen la misma cantidad de observaciones en cada una de las entrevistas.

```
panel_completo<- panel %>% filter(total_n==5)

panel_completo %>%
  tabyl(n_ent)

## #> #> #> n_ent      n    percent
## #> #> 1 59217      0.2
## #> #> 2 59217      0.2
## #> #> 3 59217      0.2
## #> #> 4 59217      0.2
## #> #> 5 59217      0.2
```

Así, ya tenemos una base que contiene la información de los individuos que han sido entrevistados cinco veces. Esta base tiene un formato “largo”. En la siguiente sección discutiremos más de los formatos para trabajar una base longitudinal.

## B. Formato largo vs. formato ancho

Como hemos señalado, la base que construimos tiene más de un renglón por cada persona. Esto es lo que consideraremos un formato *long* o largo. Tal como se muestra en el Cuadro VII-1, cada individuo tiene

un número de entrevista y las variables pueden cambiar a lo largo de esa persona.

**Cuadro VII-1. Ejemplo de cómo se ve una persona en la base de datos con formato largo**

<b>id_persona</b>	<b>n_ent</b>	<b>“clase1”</b>
Personal	1	PEA
Personal	2	PEA
Personal	3	PEA
Personal	4	PNEA
Personal	5	PNEA

Otro formato que puede ser útil para el análisis es el formato ancho. En este formato cada línea es una persona y no existe variable de número de entrevista, pero sí tenemos cada una de nuestras variables repetidas de acuerdo con el número de la observación, tal y como muestra el Cuadro VII-2. Visualmente, un formato es más ancho que el otro; en términos de sus dimensiones (número de observaciones y variables), uno es más ancho y el otro más largo. De ahí los nombres que usamos para referirnos a ellos.

**Cuadro VII-2. Ejemplo de cómo se ve una persona en la base de datos con formato ancho**

<b>id_persona</b>	<b>“clase1”</b>	<b>“clase2”</b>	<b>“clase3”</b>	<b>“clase4”</b>	<b>“clase5”</b>
Personal	PEA	PEA	PNEA	PNEA	PNEA

Vamos a hacer una transformación de nuestra base a formato ancho. Anteriormente hicimos alguna de estas transformaciones para graficar en “ggplot()”. En esta ocasión lo haremos para observar nuestra base de datos.

Primero hacemos una base original mucho más pequeña, con menos variables, para poder observar claramente cómo funciona este comando.

```

panel_long<-panel_completo %>%
  select(cd_a, ent, con, v_sel,n_hog, h_mud, n_ren, sex, eda, clase1,
         clase2, n_ent)

panel_wide <- panel_long %>%
  pivot_wider(id_cols=c(cd_a, ent, con, v_sel,n_hog, h_mud, n_ren),
             names_from=n_ent, values_from = c(clase1, clase2))

names(panel_wide)

## [1] "cd_a"      "ent"       "con"       "v_sel"     "n_hog"    "h_mud"
## [7] "n_ren"     "clase1_1"  "clase1_2"  "clase1_3"  "clase1_4"  "clase1_5"
## [13] "clase2_1"   "clase2_2"   "clase2_3"   "clase2_4"   "clase2_5"

```

¿Qué se puede observar?

- Las observaciones coinciden con los tabulados que hicimos por “n\_ent” anteriormente.
- Las variables “sex” y “eda” no se incluyeron en la nueva base ancha. Tenemos siempre que establecer si es una variable identificadora de nuestra unidad individual en la opción “id\_cols” o si proveerá los nombres o valores. Como estas variables las dejamos fuera de estas opciones, no están en la nueva base.

Ambas configuraciones son útiles para analizar la base de datos; su uso muchas veces depende del gusto del investigador(a) y con cuál de las dos se siente más cómodo(a). Para hacer el análisis descriptivo, usualmente las personas prefieren la configuración “ancha”. Por ejemplo, podemos observar rápidamente cómo se comportan los individuos a lo largo del panel en términos de su ocupación.

Si queremos observar cómo están ocupados quienes responden la primera entrevista y cómo se comportan a lo largo de los trimestres:

```

#Tabulado con formato ancho
panel_wide %>%
  filter(clase2_1==1) %>% #filtra por ocupados en el primer trimestre
  mutate(clase2_5=as_label(clase2_5)) %>% #para las etiquetas
  tabyl(clase2_5, show_missing_levels = FALSE) # tabulados para el último trimestre

##          clase2_5      n    percent
## Población ocupada 22339 0.84041232
## Población desocupada 513 0.01929950
##           Disponibles  670 0.02520597
## No disponibles    3059 0.11508220

```

Este tabulado nos dice que, en México, para el trimestre IV de 2018, el 84% de quienes estaban ocupados también lo estaban durante el trimestre IV de 2019.

Es importante señalar que los factores de expansión de este panel ya no serán los mismos con respecto a la muestra completa. Nos hemos quedado con alrededor de la quinta parte de las personas que se siguen a lo largo de cinco trimestres. Además, hemos perdido una parte importante por la atrición.

Estos datos son importantes y nos muestran una basta riqueza al analizar a los individuos en diferentes momentos en el tiempo. En la siguiente sección revisaremos una técnica específica que nos permite visualizar y analizar las trayectorias de las variables laborales a lo largo de los cinco trimestres.

## C. El análisis de secuencias. Una introducción

Las secuencias se pueden definir como una lista ordenada de estados, por ejemplo (ocupado/desocupado), o bien, definición de eventos (tiene un hijo, en unión).

Específicamente con la ENOE, podemos hacer un tipo de secuencias que se llama de “estados” —esto es así porque asumimos que lo reportado en un trimestre se mantiene a lo largo del tiempo que dura— y colocamos los estados unos detrás de otros hasta completar una secuencia de cinco estados.

```
if (!require("pacman")) install.packages("pacman") # instala pacman si se requiere
## Loading required package: pacman

#cargan Los paquetes necesarios para La práctica de este capítulo
pacman::p_load(tidyverse, haven, janitor, extdplyr, TraMineR, WeightedCluster, wesanderson,
sjlabelled)

rm(list=ls()) #botar objetos
load("panel_completo.RData") # se descarga este archivo en La carpeta de trabajo que has establecido
ls() # tenemos nuestra base de datos que armamos en el capítulo anterior.

## [1] "panel"           "panel_completo"
```

En algunas ocasiones será muy importante crear un identificador de individuo, por lo que en este comando creamos uno para la base en formato largo que estamos trabajando.

```
panel_completo<- panel_completo %>%
  group_by(cd_a, ent, con, v_sel,n_hog, h_mud, n_ren) %>%
    mutate(id = cur_group_id()) %>%
  ungroup()
```

A continuación seleccionamos un subconjunto de variables. Esta base está acomodada con menos variables, pues trabajaremos únicamente con la población en edad de trabajar. Para facilitar el trabajo y los objetos gráficos en términos de cómputo, utilizaremos nuestros identificadores recién creados. Vamos a usar tanto el formato largo el ancho:

```
panel_long<-panel_completo %>%
  filter(edad>14 & edad<98) %>% #filtro de edad a trabajar
  select(id, sex, edad, clase1, clase2, n_ent) %>% # variables
  group_by(id) %>%
    add_tally %>% # un nuevo conteo de casos, al borrar edad creamos casos
incompletos
  filter(n==5) %>% # nos quedamos con quienes tengan los cinco estados
  ungroup() # para volver a la base desagrupada

panel_wide <- panel_long %>%
  pivot_wider(id_cols=c(id, sex),
             names_from=n_ent, values_from = c(edad,clase1, clase2)) #cambia
a formato ancho
```

Para una parte del análisis será importante tener vectores de carácter para nuestras etiquetas y una selección de las variables en análisis. Se trabajará con los estados de inserción en el mercado de trabajo definidos por cuatro estados: población ocupada, población desocupada, población disponible, población no disponible para trabajar; esto de acuerdo a la variable precodificada de la ENOE “clase2”. Esta variable ya se discutió en el capítulo III, específicamente en la Figura III-2.

A continuación, se encuentra el código de los vectores con etiquetas:

```
clase2.lab <- get_labels(panel$clase2)[-1] # un vector con las etiquetas, menos la primera
que es "No Aplica"
clase2.lab
## [1] "Población ocupada"      "Población desocupada" "Disponibles"
## [4] "No disponibles"

clase2.shortlab <- c("OCU", "DES", "DIS", "NOD") #Vector de etiquetas cortas.

var_seq<-c("clase2_1", "clase2_2", "clase2_3", "clase2_4", "clase2_5") # variables con
las que se ejecutarán las secuencias

n_ent.shortlab <- c("1a ent.", "2da ent", "3ra ent", "4ta ent", "5ta ent") #Vector de
etiquetas cortas
```

## a. Gráficas aluviales

Una manera de evidenciar cómo los individuos cambian entre los estados de cada trimestre de su entrevista son las gráficas aluviales. Estas gráficas, también conocidas como de hilos, son un tipo de Diagrama de Sankey que sirven para representar flujos de datos y su magnitud. En este caso, un flujo es una trayectoria específica y lo ancho de su flujo es la frecuencia de ésta.

Para elaborar esta gráfica se necesita la paquetería *easyalluvial*. Por ser un paquete en desarrollo se debe instalar desde el código fuente que el desarrollador publica en el sitio github; por ello, lo instalaremos de manera ligeramente diferente y no a través de *pacman*.

```
remotes::install_github("erblast/easyalluvial")
library(easyalluvial) #carga la librería
```

*Easyalluvial* es un paquete basado en ggplot y compatible con él. Para optimizar la visualización de la gráfica haremos un par de arreglos. Este paquete está optimizado para que nuestras variables sean de tipo cadena o factor y no “haven\_labelled”, como las importamos. Del mismo modo, las etiquetas de los estados son bastante largas, por lo que utilizaremos las etiquetas cortas que guardamos en nuestro objeto “clase2.shortlab”.

```
panel_long$clase2<-set_labels(panel_long$clase2, labels=clase2.shortlab) # establece
las etiquetas
panel_long$n_ent<-set_labels(panel_long$n_ent, labels=n_ent.shortlab) # establece
las etiquetas

panel_alluvial<- panel_long %>%
  select(c(clase2, n_ent, id)) %>% #nos quedamos sólo con las variables de la secuencia,
id y n_ent
  mutate_at(vars(clase2, n_ent), as_label) %>% # cambian a sus valores de etiqueta
  mutate(id=as.factor(id)) # cambia de numérico a factor

glimpse(panel_alluvial) #revisamos los cambios

## Observations: 222,565
## Variables: 3
## $ clase2 <fct> OCU, OCU, OCU, OCU, OCU, OCU, OCU, NOD, NOD, OCU, OCU...
## $ n_ent <fct> 1a ent., 2da ent., 3ra ent., 4ta ent., 5ta ent., 1a ent., 1a ...
## $ id <fct> 1, 1, 1, 1, 1, 2, 3, 4, 2, 3, 4, 2, 3, 4, 2, 3, ...
```

Para que nuestros colores sean vistosos, seguiremos utilizando la paleta de colores de las películas de Wes Anderson. Como son

cuatro categorías, creamos una paleta de cuatro colores, de una escala cualitativa o discreta:

```
cols <- wes_palette(n=4, name="Darjeeling2", type="discrete") # un vector de colores
creados con la paleta de Wes Anderson
cols
```

**Figura VII-2. Paleta de colores de la película *Viaje a Darjeeling***



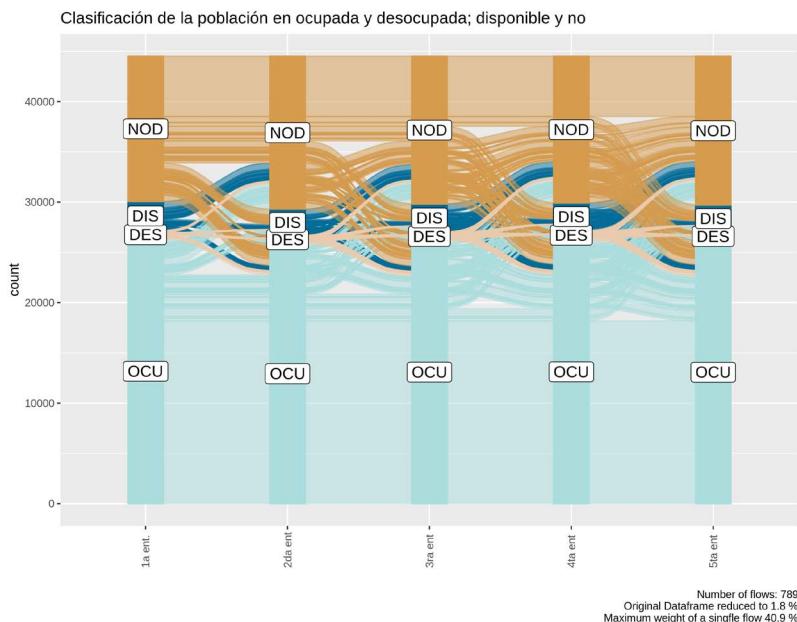
Para realizar la gráfica aluvial, se utiliza el panel en formato largo para la población de 12 años y más. Seleccionamos únicamente las variables de secuencia de la variable “clase2”. El comando es “alluvial\_long”.

```
allu<-alluvial_long(data=panel_alluvial, #selección del objeto de base de datos
key = n_ent, # Variable que identifica los estados
value = clase2, # Variable con la que estableceremos los cambios
id= id, # identificador único de los individuos
fill_by='value', # Define los flujos en términos de la variable
declarada en "value"
col_vector_flow = cols, # Color para los flujos
col_vector_value = cols) # Color para los recuadros

## Warning: Column `n_ent` has different attributes on LHS and RHS of join

allu + ggtitle(paste(get_label(panel_alluvial$clase2))) # Lo imprime con el título de
la etiqueta de la variable como título
```

### Gráfica VII-1. Diagrama de hilos o aluvial de la población en edad de trabajar



FUENTE: elaboración propia con base en la ENOE.

Con esta gráfica se muestra cómo los trabajadores cambian de estado en períodos de observación tan pequeños como trimestres. Vemos que hay una gran proporción de personas que se mantienen ocupadas los cinco trimestres y otro gran grupo se queda en la no disponibilidad. Aunque se aprecian algunos cambios a lo largo de los cinco trimestres, no queda duda de que se necesitan más herramientas de análisis. Con ello, se procede en el siguiente acápite.

### b. Análisis de secuencias con el paquete *TraMineR*

El paquete *TraMineR* es fundamental para el análisis de secuencias. Genera análisis descriptivos, formas de agrupamiento y medidas que analizan específicamente cómo se comportan esas secuencias. Aquí nos vamos a concentrar en el análisis de secuencias

de estado, tal como se describe en Gabadinho, Ritschard, Müller y Studer (2011).

Primero vamos a definir nuestra secuencia con el comando “seqdef()”. Usaremos nuestra base en formato ancho. Este formato ancho es compatible con el formato “STate-Sequence (STS)” del paquete; es decir, los estados sucesivos representan una posición o columna en nuestra base de datos en formato ancho. La definición de la secuencia la guardaremos en un nuevo objeto.

```
clase2.seq <- seqdef(data=panel_wide, # base de datos
                      var=var_seq, # variables en análisis
                      states = clase2.shortlab, #estados, preferimos usar etiquetas cortas
                      xtstep = 1) #Número de marcas en el eje x para gráficas
## [>] state coding:
##      [alphabet] [label] [long label]
##      1          OCU      OCU
##      2          DES      DES
##      3          DIS      DIS
##      4          NOD      NOD
## [>] 44513 sequences in the data set
## [>] min/max sequence length: 5/5
```

### c. Descripción de las secuencias

Para describir las secuencias podemos tabularlas y graficarlas. Para establecer qué frecuencias son más comunes, usamos el comando “seqtab()”:

```
seqtab(clase2.seq, #objeto de secuencia
       idxs=1:10) # establece el número de secuencias a tabular, si no se establece se tabulan todas
##           Freq Percent
## OCU/5        18193   40.9
## NOD/5        5986    13.4
## OCU/4-NOD/1    786    1.8
## NOD/1-OCU/4    675    1.5
## OCU/1-NOD/1-OCU/3  664    1.5
## DIS/1-NOD/4    635    1.4
## OCU/3-NOD/1-OCU/1  631    1.4
## NOD/2-DIS/1-NOD/2  610    1.4
## NOD/4-DIS/1      596    1.3
## NOD/3-DIS/1-NOD/1  595    1.3
```

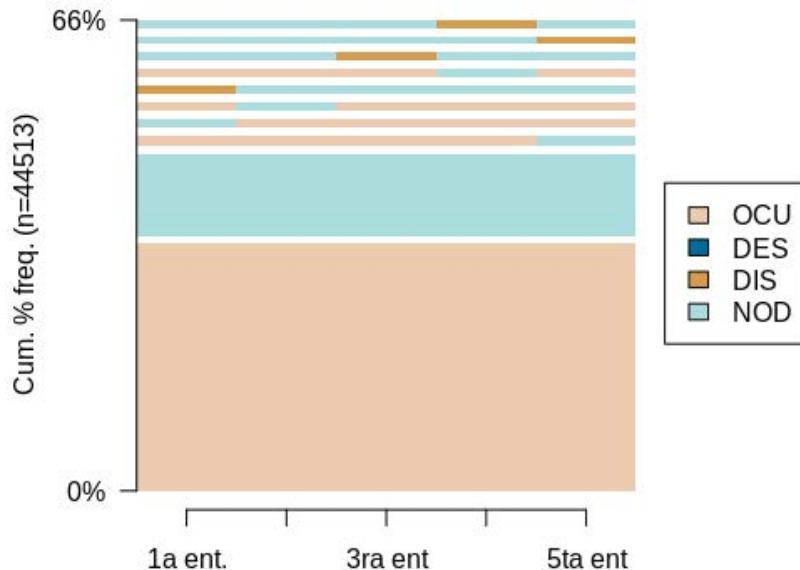
Como observamos, el 40.85% de los trabajadores tiene una secuencia donde los cinco estados son de ocupación; mientras que el 13.4% pasa en estado de no disponibilidad los cinco trimestres. Esto nos ayuda a comprender de mejor manera lo que se observa en la Gráfica VII-1.

El nombre de la secuencia de este objeto establece el estado descrito en su versión corta “OCU”, seguido de una diagonal, “/”, y el número subsiguiente es el número de trimestres en la condición.

En este paquete también podemos hacer unos gráficos que representan estas secuencias, donde además se ordenan de mayor a menor de acuerdo con su frecuencia:

```
seqfplot(seqdata=clase2.seq, # objeto de la definición de secuencias  
        with.legend = "right", # posición de la leyenda  
        border = NA, # opciones de bordes  
        cpal=cols, #colores de nuestro vector  
        xlab=n_ent.shortlab) # Para etiquetar el eje x con el número de entrevistas
```

**Gráfica VII-2. Frecuencia de las secuencias de inserción laboral.  
México, panel que inicia el trimestre IV de 2018**

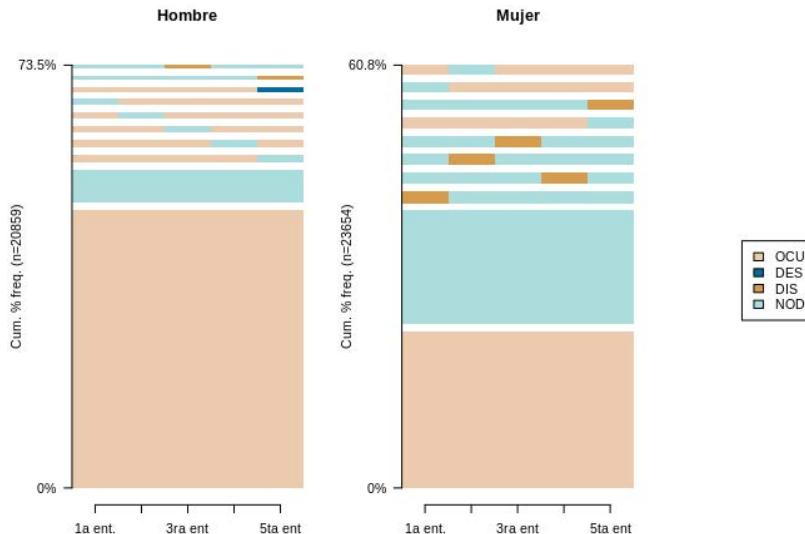


FUENTE: elaboración propia con base en la ENOE.

Podemos incluir alguna variable cualitativa que sea relevante para el análisis agregando una línea más al código. En este caso, se presenta con la condición de ser hombre o mujer:

```
seqfplot(seqdata=clase2.seq, # objeto de la definición de secuencias
         with.legend = "right", # posición de la leyenda
         border = NA, # opciones de bordes
         cpal=cols, #colores de nuestro vector
         xlab=n_ent.shortlab,
         group=as_label(panel_wide$sex) ) #Se realiza para cada categoría
```

**Gráfica VII-3. Frecuencia de las secuencias de inserción laboral, según sexo. México, panel que inicia el trimestre IV de 2018**



La Gráfica VII-3 presenta claramente cómo las trayectorias de corto plazo son diferentes para hombres y mujeres.

Otra medida importante es el tiempo promedio en cada estado. Como nuestras observaciones son por trimestre, estos promedios tienen como unidad un trimestre. El cálculo se realiza con el comando “seqmean()”:

```
seqmean(seqdata=clase2.seq)

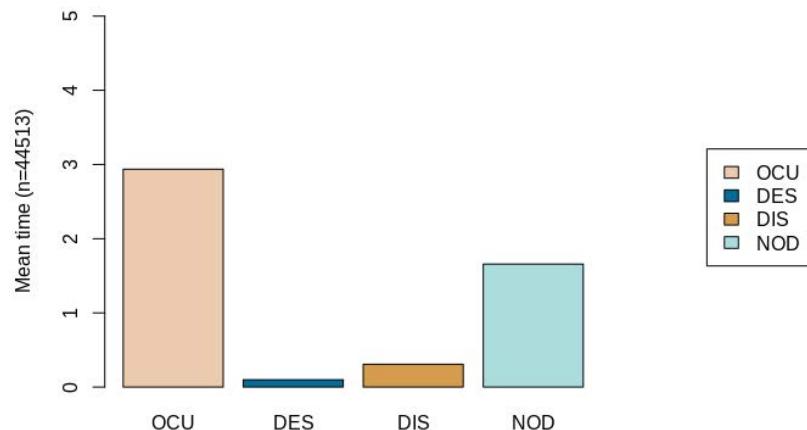
##      Mean
##  OCU 2.937
##  DES 0.099
##  DIS 0.307
##  NOD 1.657
```

Como vemos, en promedio los trabajadores pasan casi tres trimestres en estados de ocupación; así como 1.65 trimestres en estados de no disponibilidad.

Otra manera de visualizar esta información es con un gráfico de tipo promedio con el comando “seqmtpplot”:

```
seqmtpplot(clase2.seq,  
           with.legend="right",  
           cpal=cols  
)
```

**Gráfica VII-4. Trimestres promedio en cada estado de (no) inserción laboral. México, panel que inicia el trimestre IV de 2018**



FUENTE: elaboración propia con base en la ENOE.

Si queremos hacerlo para grupos, por ejemplo, para evidenciar las diferencias por sexo, podemos utilizar los siguientes comandos:

```
by(clase2.seq,  
   as_label(panel_wide$sex),  
   seqmean)  
  
## as_label(panel_wide$sex): Hombre  
##      Mean  
##  OCU 3.70  
##  DES 0.13  
##  DIS 0.22  
##  NOD 0.96  
## -----
```

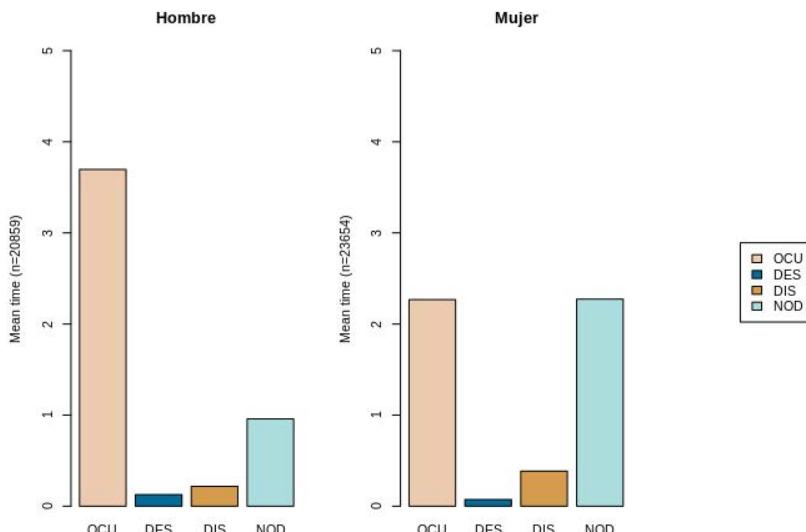
```
## as_label(panel_wide$sex): Mujer
##      Mean
## OCU 2.267
## DES 0.074
## DIS 0.386
## NOD 2.273
```

Las diferencias por sexo son bastante grandes y dan cuenta de las distinciones en las dinámicas entre hombres y mujeres en el mercado de trabajo y la división sexogenérica entre trabajo no remunerado y remunerado. De ahí que las mujeres tengan más trimestres en la no disponibilidad que los hombres.

También se puede graficar esta información introduciendo el argumento “group” en la función “seqmtpplot()”.

```
seqmtpplot(clase2.seq,
           with.legend="right",
           cpal=cols,
           group=as_label(panel_wide$sex))
```

**Gráfica VII-5. Trimestres promedio en cada estado de (no) inserción laboral, según sexo. México, panel que inicia el trimestre IV de 2018**



FUENTE: elaboración propia con base en la ENOE.

## D. Tasas de transición

Otra manera de analizar la información longitudinal es con las tasas de transición entre los estados. Una tasa de transición mide la probabilidad de cambiar de un estado  $s_i$  a un estado  $s_j$ . De acuerdo a Gaba-dinho *et al.* (2011), esta probabilidad se denota por una probabilidad condicional, como se lee a continuación:

$$p(s_j|s_i) = \frac{\sum_{t=1}^{L-1} n_{t,t+1}(s_i, s_j)}{\sum_{t=1}^{L-1} n_t(s_i)}$$

Donde  $L$  es el largo máximo de la secuencia, es decir, el número máximo de estados observados;  $n_t(s_i)$  es el número de secuencias que en la observación o posición  $t$  tiene el estado  $s_i$ ; mientras que  $n_{t,t+1}(\cdot, s_j)$  son las secuencias con el estado  $s_i$  en la posición  $t$  y observación  $t+1$  tienen el estado  $s_j$ .

Reescribamos esta ecuación para nuestros casos con la ENOE, utilizando la variable “clase2”. Pensemos que  $s_i$  es “ocupación” (ocu) y  $s_j$  es “desocupación” (des). Calculamos la transición al desempleo dado que se está ocupado.

$$p(s_{des}|s_{ocu}) = \frac{\sum_{t=1}^4 n_{t,t+1}(s_{ocu}, s_{des})}{\sum_{t=1}^4 n_t(s_{ocu})}$$

$$p(s_{des}|s_{ocu}) = \frac{n_{ocu(t_1)->des(t_2)} + n_{ocu(t_2)->des(t_3)} + n_{ocu(t_3)->des(t_4)} + n_{ocu(t_4)->des(t_5)}}{n_{ocu(t_1)} + n_{ocu(t_2)} + n_{ocu(t_3)} + n_{ocu(t_4)}}$$

Es decir, es una probabilidad porque en el denominador se tiene a todos los que están en riesgo de experimentar la transición, mientras que en el numerador a quienes la experimentaron y cuántos pasaron de ocupados a desempleados.

Con un par de tabulados en nuestra base en formato ancho, podemos terminar de escribir y sustituir en la ecuación cada uno de los elementos.

Calculando  $n_{ocu(t_1)} \rightarrow des(t_2)$

```
#t=1 y t+1=2
t1t2<-panel_wide %>% filter(clase2_1==1 & clase2_2==2) %>% tally()
t1t2
```

```
## # A tibble: 1 × 1
##      n
##  <int>
## 1    436
```

Calculando  $n_{ocu(t_2)} \rightarrow des(t_3)$

```
#t=2 y t+1=3
t2t3<-panel_wide %>% filter(clase2_2==1 & clase2_3==2) %>% tally()
t2t3

## # A tibble: 1 × 1
##      n
##  <int>
## 1    409
```

Calculando  $n_{ocu(t_3)} \rightarrow des(t_4)$

```
#t=3 y t+1=4
t3t4<-panel_wide %>% filter(clase2_3==1 & clase2_4==2) %>% tally()
t3t4

## # A tibble: 1 × 1
##      n
##  <int>
## 1    461
```

Calculando  $n_{ocu(t_4)} \rightarrow des(t_5)$

```
#t=4 y t+1=5
t4t5<-panel_wide %>% filter(clase2_4==1 & clase2_5==2) %>% tally()
t4t5

## # A tibble: 1 × 1
##      n
##  <int>
## 1    465
```

Esto significa la suma de  $436 + 409 + 461 + 465 = 1771$ . Este sería el numerador que incluye a todos los ocupados que en el siguiente periodo llegaron a desocuparse. Para el denominador, necesitamos todos los que estaban ocupados del periodo 1 al 4. Esto es más fácil calcularlo con la base en formato largo.

Calculando  $n_{ocu(t_1)} + n_{ocu(t_2)} + n_{ocu(t_3)} + n_{ocu(t_4)}$

```
ocut1_t4<-panel_long %>% filter(n_ent!=5 & clase2==1) %>% tally()
ocut1_t4

## # A tibble: 1 × 1
```

```
##      n
## <int>
## 1 104585
```

Finalmente, nuestra tasa de transición es  $p(s_{des}|s_{ocu}) = 0.0169336$ .

```
(t1t2+t2t3+t3t4+t4t5)/ocut1_t4
##      n
## 1 0.01693359
```

Así hacemos todas las combinaciones posibles de transición entre los cuatro estados y obtenemos 24 transiciones. El paquete las calcula con la función “seqrate()”.

```
clase2.trate<-seqrate(clase2.seq) #guardamos las tasas de transición en un objeto
## [>] computing transition probabilities for states OCU/DES/DIS/NOD ...
round(clase2.trate,2) # Las imprimimos a dos decimales
##           [-> OCU] [-> DES] [-> DIS] [-> NOD]
## [OCU ->]    0.87    0.02    0.02    0.09
## [DES ->]    0.52    0.19    0.08    0.21
## [DIS ->]    0.20    0.03    0.18    0.59
## [NOD ->]    0.16    0.01    0.11    0.72
```

Si revisamos el valor de la primera fila y de la segunda columna, es justo el valor que calculamos a “mano”. Realmente, esta función nos ahorra muchas operaciones, incluso podemos imprimirlo sin redondear los dígitos:

```
clase2.trate[1,2]
## [1] 0.01693359
```

Revisemos un poco estos resultados. La diagonal muestra las “transiciones” hacia un mismo estado, en realidad se trata de permanencias. Si estos valores son muy cercanos a 1, quiere decir que hay poco movimiento, pero notamos que los valores altos son los estados de ocupación y los de no disponibilidad. En cambio, sólo el 19% de la población desempleada se quedaría como tal entre un trimestre y otro. Ello da cuenta de que el desempleo tiene una dinámica muy amplia, al igual que la condición de disponibilidad. Más del 50% de las personas que están desempleadas pasan a un estado de ocupación en el siguiente trimestre. Sin duda, ello indicaría que el mercado laboral las está absorbiendo. Lo que aquí se debe recordar es lo aprendido

en capítulos anteriores: el mercado laboral mexicano también tiene condiciones precarias y, además, pocas personas se pueden quedar mucho tiempo en condición de búsqueda.

Ahora, estas transiciones son para toda la población. Muchas transiciones son heterogéneas para grupos distintos. Por ejemplo, para tener más claras las diferencias entre sexos, replicamos esto según sexo:

```
bysex_seqrate<-by(clase2.seq,
  as_label(panel_wide$sex),
  seqrate)

##  [] computing transition probabilities for states OCU/DES/DIS/NOD ...
##  [] computing transition probabilities for states OCU/DES/DIS/NOD ...

#Hombres
round(bysex_seqrate$Hombre,2)

##          [-> OCU] [-> DES] [-> DIS] [-> NOD]
## [OCU ->]      0.91     0.02     0.01     0.06
## [DES ->]      0.57     0.21     0.07     0.15
## [DIS ->]      0.26     0.05     0.20     0.49
## [NOD ->]      0.21     0.02     0.11     0.66

#Mujeres
round(bysex_seqrate$Mujer,2)

##          [-> OCU] [-> DES] [-> DIS] [-> NOD]
## [OCU ->]      0.81     0.01     0.03     0.14
## [DES ->]      0.43     0.16     0.10     0.32
## [DIS ->]      0.17     0.02     0.17     0.64
## [NOD ->]      0.14     0.01     0.11     0.74
```

Se observa cómo las mujeres tienen tasas de transición más bajas hacia la ocupación, independientemente de los estados de inicio, con respecto a los hombres. También, las mujeres tienen tasas de transición hacia la no disponibilidad mucho más altas que los hombres.

Sin duda, el análisis de secuencias es muy rico en sus aplicaciones. Aquí sólo tratamos un par de elementos. Actualmente, en la Asociación de Análisis de Secuencias (<https://sequenceanalysis.org/>) se da cabida a los nuevos aportes de este análisis a las ciencias sociales. Los aportes van desde incluir los grandes datos, análisis de redes y más.

Estas breves aplicaciones para el caso mexicano nos han permitido ver que estar en la ocupación no necesariamente implica que sea un empleo que se mantenga en el tiempo, incluso al observar una

ventana tan corta como quince meses. Las frecuencias y gráficas de las secuencias (tanto aluviales como de frecuencia relativa), nos permitieron descubrir que quienes inician la observación longitudinal en ocupación no se quedan en esa condición y que, incluso, pueden transitar hasta la no disponibilidad. Del mismo modo, pudimos evidenciar que las secuencias de inserción de la población en edad de trabajar están diferenciadas entre hombres y mujeres. Las mujeres tienen trayectorias de corto plazo más intermitentes que los hombres, con muchas más entradas y salidas; esto se debe de analizar a la luz de los roles de género y cómo las mujeres también realizan trabajo no remunerado dentro de los hogares.

Los campos del uso del panel de la ENOE son amplios y no se circunscriben únicamente al análisis de secuencias. Tanto la construcción del panel como la aplicación de varios paquetes para el análisis de secuencias son una barrera para este interesante análisis que nos permite observar el dinamismo de las personas en los mercados de trabajo en períodos de quince meses. Pero sin duda, es una veta que los y las estudiosas del mercado de trabajo debemos profundizar.

## Palabras finales

El camino para estudiar un problema es único para quien investiga. Este libro ha intentado abonar en hacer un menú más amplio y accesible para quien desee estudiar los mercados de trabajo mexicanos. En ningún momento este libro es un recetario, es más, el objetivo es dar mayores herramientas para que la creatividad se expanda al buscar nuevos caminos y preguntas de investigación.

Méjico ha destacado por su amplia disponibilidad de información. Esta información se recoge, en parte, con los ingresos públicos que provienen de los impuestos ciudadanos. Explotar la información es, además de un derecho, un deber que deberíamos de considerar los y las científicas sociales.

No obstante, existen muchos obstáculos para esta explotación. A pesar de que hoy *R* es un programa libre, no es aún tan utilizado en las ciencias sociales como otros paquetes que son bastante costosos. Ello porque estos programas presentaron plataformas que eran más fáciles de interactuar con el usuario. *R* ha tenido varias interfaces que han ido acortando la brecha con esas plataformas, *RStudio* es sin duda una interfaz que está siendo cada día más utilizada.

A pesar de este aumento en el uso de *R*, existe otra brecha que este libro ha querido aportar: en general, la documentación está escrita en inglés. Si bien la comunidad de *R* es cada vez más grande y activa (véase <<https://latin-r.com/>>), sin duda hay mucho camino por recorrer.

Por otro lado, también existe una brecha importante para poder enseñar a investigar desde una mirada cuantitativa. Los ejemplos con los que se construyen los manuales muchas veces interpelan poco al usuario. En este caso, se ha optado por desarrollar códigos con una fuente específica y de gran trayectoria, como lo es la Encuesta Nacional de Ocupación y Empleo. Cada tabulado, gráfica o resultado de un modelo presentados en este libro proviene de la realidad mexicana.

Del mismo modo, al presentar una fuente de información “real”, este libro ha intentado solventar problemas a los que uno se enfrenta

en el análisis de datos: etiquetar, cambiar de formato, fusionar información. Muchos manuales se realizan con bases de datos muy bien comportadas y demasiado *ad hoc* a las herramientas que se exponen. Esto rara vez es la realidad del proceso al que se enfrenta un investigador o investigadora social, por lo que a lo largo del libro también se han incluido pequeños ejemplos de estos elementos.

No se han podido abarcar todas las herramientas necesarias para el análisis de los mercados de trabajo. Por ejemplo, pueden parecer muy interesantes los métodos específicos para datos categóricos (i.e. regresión logística, regresión logística multivariada), o métodos multivariados para resumir información o variables (análisis de clústers, análisis factorial, componentes principales), métodos para análisis diferentes unidades de información (análisis jerárquico multinivel), entre muchos otros más; no obstante, este libro intentó dar los suficientes elementos para lo básico de un modelo estadístico y que su aplicación se pueda llevar a cabo. Y, de ahí, lograr más autonomía para quien quiera exceder de los ámbitos presentados en este libro.

La Encuesta Nacional de Ocupación y Empleo, como revisamos en el capítulo 1, considera elementos de diferentes grupos y se apega a varios criterios internacionales. Sin duda, es una encuesta que deberá cambiar pronto ante las nuevas disposiciones de las Conferencias Internacionales de Estadísticos del Trabajo (CIET) y exhortamos a mantenerse al día de los cambios que puedan ocurrir. Algunos de estos elementos ya han sido revisados y podrían tener cambios fundamentales en algunos elementos conceptuales y metodológicos (véase Padrón, Gandini & Navarrete, 2017, para un detalle en diferentes variables y conceptos).

Actualmente, y como respuesta a la pandemia, el INEGI estableció una encuesta telefónica en los meses de abril a junio: la Encuesta Telefónica de Ocupación y Empleo (ETOE), siguiendo el cuestionario ampliado de la ENOE. Posteriormente, INEGI presentó una nueva enoe, la ENOE-Nueva Construcción (ENOEN<sup>N</sup>), que incluye tanto entrevistas presenciales como telefónicas. De nuevo, los cuestionarios se han mantenido con cambios en los identificadores de las tablas de las bases de datos. Mucho del análisis expuesto en este libro se puede mantener para estas nuevas fuentes de información sobre los mercados laborales mexicanos.

Del mismo modo, exhortamos a mantenerse al día en la comunidad de *R*. Algunos de los códigos utilizados en el presente libro funcionan para el momento en que fue escrito. Los comandos y paquetes están vivos gracias a la contribución de todos los usuarios. Es probable que aparezcan funciones más sencillas, o bien, que alguna sintaxis de las funciones haya cambiado.

## Referencias

- Abramo, L. & Valenzuela, M. E. (2005). Women's labour force participation rates in Latin America. *International Labour Review*, 144(4), 369-400. <https://doi.org/10.1111/j.1564-913X.2005.tb00574.x>
- Allaire, J. J., Xie [aut, cre], Y., McPherson, J., Luraschi, J., Ushey, K. et al. (2020). rmarkdown: Dynamic Documents for R (Versión 2.1). Recuperado de <https://CRAN.R-project.org/package=rmarkdown>
- Amrhein, V., Greenland, S. & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 567(7748), 305-307. <https://doi.org/10.1038/d41586-019-00857-9>
- Angrist, J. D. & Pischke, J. S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.
- Behrendt, S. (2014). lm.beta: Add Standardized Regression Coefficients to lm-Objects (Versión 1.5-1). Recuperado de <https://CRAN.R-project.org/package=lm.beta>
- Binelli, C. & Rubio, M. (2013). The returns to private education: Evidence from Mexico. *Economics of Education Review*. Recuperado de <https://www.sciencedirect.com/science/article/pii/S027277571300085X>
- Buuren, S. van, Groothuis, K., Vink, G., Schouten, R., Robitzsch, A., Doove, L., ... Gray, B. (2020). mice: Multivariate Imputation by Chained Equations (Versión 3.8.0). Recuperado de <https://CRAN.R-project.org/package=mice>
- Caamal, C. G. (2017). Rendimientos decrecientes de la escolaridad en México. *Estudios Económicos (Méjico, D.F.)*, 32(1). Recuperado de [http://www.scielo.org.mx/scielo.php?pid=S0186-72022017000100027&script=sci\\_arttext&tlang=pt](http://www.scielo.org.mx/scielo.php?pid=S0186-72022017000100027&script=sci_arttext&tlang=pt)
- Campos, R. M. (2013). Efectos de los ingresos no reportados en el nivel y tendencia de la pobreza laboral en México. *Ensayos Revista de Economía*, XXXII(2), 23-54.

- Cendejas, J. N. (2017). Higher education and work: Towards the construction of an information system on college graduates. *Diálogos sobre educación*. Recuperado de <https://www.redalyc.org/jatsRepo/5534/553458101012/html/index.html>
- Chen, J. (2020). gglorenz: Plotting Lorenz Curve with the Blessing of “ggplot2” (Versión 0.0.2). Recuperado de <https://CRAN.R-project.org/package=gglorenz>
- Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL) (2020). Líneas de pobreza por ingresos. Recuperado el 19 de marzo de 2020, de <http://sistemas.coneval.org.mx/InfoPobreza/Pages/wfrLineaBienestar?pAnioInicio=2016&pTipoIndicador=0>
- Cortés, F. & Rubalcava, R. M. (1993). Consideraciones sobre el uso de la estadística en las ciencias sociales: Estar a la moda o pensar un poco. En I. Méndez & P. González Casanova (Eds.), *Matemáticas y ciencias sociales* (pp. 227-267). Distrito Federal: unam-ccih; Grupo Editorial Miguel Angel Porrúa.
- Cranmer, S., Gill, J., Jackson, N., Murr, A. & Armstrong, D. (2020). hot.deck: Multiple Hot-Deck Imputation (Versión 1.1-1). Recuperado de <https://CRAN.R-project.org/package=hot.deck>
- Cruz, L. M. S. & Reyes, P. R. (2015). Persistencias y cambios en la participación laboral en el Estado de México durante el periodo 2000-2012. *Papeles de población*, 21(83). Recuperado de <https://dialnet.unirioja.es/servlet/articulo?codigo=5297203>
- Dorfman, R. (1979). A Formula for the Gini Coefficient. *The Review of Economics and Statistics*, 61(1), 146-149. JSTOR. <https://doi.org/10.2307/1924845>
- Ellis, G. F., Lumley, T., Źółtak, T. & Schneider, B. (2020). srvyr: ‘dplyr’-Like Syntax for Summary Statistics of Survey Data (Versión 0.3.7). Recuperado de <https://CRAN.R-project.org/package=srvyr>
- Escoto, A. (2020). La inserción laboral de las mujeres en México: Una mirada longitudinal de corto plazo. *Coyuntura Demográfica. Revista sobre los Procesos Demográficos en México Hoy*, 18(julio), 59-67.
- Escoto, A., Márquez, M. C. & Prieto, V. (2020). La sobreeducación en México: ¿promotora o inhibidora de la exclusión laboral? *Revista Latinoamericana de Población*, 14(27), 115-148.

- Firke, S., Denney, B., Haid, C., Knight, R. & Grosser, M. (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data (Versión 1.2.1). Recuperado de <https://CRAN.R-project.org/package=janitor>
- Flores, J. S. R., Zamora, S. & Contreras, E. (2013). Transiciones entre el trabajo formal e informal y medios de intermediación laboral en México 2005-2010. *Banco Interamericano de Desarrollo*. Recuperado de [https://pdfs.semanticscholar.org/8287/580e2856acac52dca91083fc355159cd2c90.pdf?\\_ga=2.148436671.925693671.1573865334-732215748.1573865334](https://pdfs.semanticscholar.org/8287/580e2856acac52dca91083fc355159cd2c90.pdf?_ga=2.148436671.925693671.1573865334-732215748.1573865334)
- Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D. et al. (2019). car: Companion to Applied Regression (Versión 3.0-6). Recuperado de <https://CRAN.R-project.org/package=car>
- Freije, S., López, G. & Rodríguez, E. (2011). *Effects of the 2008-09 economic crisis on labor markets in Mexico*. elibrary.worldbank.org. Recuperado de <https://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-5840>
- Gabadinho, A., Ritschard, G., Müller, N. S. & Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, 40(1), 1-37. <https://doi.org/10.18637/jss.v040.i04>
- Gabadinho, A., Studer, M., Müller, N., Bürgin, R., Fonta, P.-A. & Ritschard, G. (2020). TraMineR: Trajectory Miner: a Toolbox for Exploring and Rendering Sequences (Versión 2.0-14). Recuperado de <https://CRAN.R-project.org/package=TraMineR>
- García, B. (2011). Las carencias laborales en México: Conceptos e indicadores. En M. E. Pacheco Gómez Muñoz, E. de la Garza Toledo & L. Reygadas (Eds.), *Trabajos atípicos y precarización del empleo* (pp. 81-113). México, D.F.: El Colegio de México.
- García, B. (2012). La precarización laboral y el desempleo en México (2000-2009). En E. de la Garza (Coord.), *La situación del trabajo en México* (pp. 91-118). México, D.F.: Plaza y Valdés.
- García, B. & Pacheco, E. (2011). La participación económica en el censo de población 2010. *Coyuntura Demográfica. Revista sobre los Procesos Demográficos en México Hoy*, 1. Recuperado de <http://www.somede.org/coyuntura-demografica/index.php>

- numero-1/item/la-participacion-economica-en-el-censo-de-poblacion-2010?highlight=WyJlZGl0aCIsInBhY2hlY28iLCJlZ-Gl0aCBwYWNoZWNVl0=
- García, R. E. (2014). *La llegada del primer hijo: Cambios en el uso del tiempo de los miembros de la pareja en México 2010-2013. Un análisis con la ENOE.* (Maestría, El Colegio de México). El Colegio de México. Recuperado de [https://colmex.userservices.exlibrisgroup.com/view/delivery/52COLMEX\\_INST/1264963100002716](https://colmex.userservices.exlibrisgroup.com/view/delivery/52COLMEX_INST/1264963100002716)
- Gelman, A., Hill, J., Su, Y.-S., Yajima, M., Pittau, M., Goodrich, B., Si, Y. & Kropko, J. (2015). mi: Missing Data Imputation and Model Checking (Versión 1.0). Recuperado de <https://CRAN.R-project.org/package=mi>
- Guzmán, F. (2009). Segregación ocupacional por género. *DemoS*, 0(015). Recuperado de <http://www.revistas.unam.mx/index.php/dms/article/view/6786/6306>
- Heckman, J. J., Lochner, L. J. & Todd, P. E. (2003). *Fifty Years of Mincer Earnings Regressions* (Working Paper Núm. 9732). National Bureau of Economic Research. <https://doi.org/10.3386/w9732>
- Heckman, J. J., Lochner, L. J. & Todd, P. E. (2006). Chapter 7 Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond. En E. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education*, 1, 307-458. Elsevier. [https://doi.org/10.1016/S1574-0692\(06\)01007-5](https://doi.org/10.1016/S1574-0692(06)01007-5)
- Henningsen, A., Toomet, O. & Petersen, S. (2019). sampleSelection: Sample Selection Models (Versión 1.2-6). Recuperado de <https://CRAN.R-project.org/package=sampleSelection>
- Hlavac, M. (2018a). oaxaca: Blinder-Oaxaca Decomposition (Versión 0.1.4). Recuperado de <https://CRAN.R-project.org/package=oaxaca>
- Hlavac, M. (2018b). stargazer: Well-Formatted Regression and Summary Statistics Tables (Versión 5.2.2). Recuperado de <https://CRAN.R-project.org/package=stargazer>
- Honaker, J., King, G. & Blackwell, M. (2019). Amelia: A Program for Missing Data (Versión 1.7.6). Recuperado de <https://CRAN.R-project.org/package=Amelia>

- Hulsizer, M. R. & Woolf, L. M. (2009). *A guide to teaching statistics: Innovations and best practices*. Malden, MA; Oxford: Wiley-Blackwell Pub.
- Instituto Nacional de Estadística y Geografía (INEGI) (2014). *La informalidad laboral. Marco conceptual y metodológico*. Recuperado de [https://www.snieg.mx/DocAcervoINN/documentacion/inf\\_nvo\\_acervo/SNIDS/ENOE/702825060459.pdf](https://www.snieg.mx/DocAcervoINN/documentacion/inf_nvo_acervo/SNIDS/ENOE/702825060459.pdf)
- INEGI (2015). *Encuesta Nacional de Ocupación y Empleo. Reconstrucción de variables. 2005 a la fecha*. Recuperado de [https://www.inegi.org.mx/contenidos/programas/enoe/15ymas/doc/recons\\_var\\_15ymas.pdf](https://www.inegi.org.mx/contenidos/programas/enoe/15ymas/doc/recons_var_15ymas.pdf)
- INEGI (2017). *Conociendo la base de datos de la ENOE. Datos ajustados a las proyecciones de población 2010*. Recuperado de [https://www.inegi.org.mx/contenidos/programas/enoe/15ymas/doc/con\\_basedatos\\_proy2010.pdf](https://www.inegi.org.mx/contenidos/programas/enoe/15ymas/doc/con_basedatos_proy2010.pdf)
- INEGI (2018). *Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH)*. Recuperado de <https://www.inegi.org.mx/programas/enigh/nc/2018/>
- INEGI (2019). *Cómo se hace la ENOE. Métodos y procedimientos*. Recuperado de [http://internet.contenidos.inegi.org.mx/contenidos/Productos/prod\\_serv/contenidos/espanol/bvinegi/productos/nueva\\_estruc/702825190613.pdf](http://internet.contenidos.inegi.org.mx/contenidos/Productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825190613.pdf)
- INEGI (2020). Encuesta Nacional de Ocupación y Empleo (ENO), población de 15 años y más de edad. Recuperado el 7 de julio de 2020, de <https://www.inegi.org.mx/programas/enoe/15ymas/>
- Koenker, R. (2020). *quantreg: Quantile Regression*. Recuperado de <https://CRAN.R-project.org/package=quantreg>
- Koneswarakantha, B. (2020). *easyalluvial: Generate Alluvial Plots with a Single Line of Code* (Versión 0.2.3). Recuperado de <https://CRAN.R-project.org/package=easyalluvial>
- Kuri Alonso, I. (2014). Mujeres y mercados de trabajo: Análisis de la segregación ocupacional por sexo en México. *International Journal of Innovation and Applied Studies*. Recuperado de <https://pdfs.semanticscholar.org/48c7/8cfb0bb4de3c2a94ebc-0311607082deba097.pdf>
- Lazarsfeld, P. (1973). De los conceptos a los índices empíricos. En R. Boudon & P. Lazarsfeld, *Metodología de las Ciencias Sociales*

- (pp. 35-46). Barcelona: Editorial Laia/Barcelona.
- Levy, S. & López, L. F. (2019). *Persistent Misallocation and the Returns to Education in Mexico*. elibrary.worldbank.org. Recuperado de <https://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-8690>
- Levy, S. & Székely, M. (2016). ¿Más escolaridad, menos informalidad? Un análisis de cohortes para México y América Latina. *El trimestre económico*, 83(332), 499-548. <https://doi.org/10.20430/ete.v83i332.232>
- Linthon, D. E. (2018). *Desajuste educativo en un mercado laboral segmentado. El caso de México, 2005-2015* (Doctorado en Ciencias Económicas, Universidad Autónoma Metropolitana). Universidad Autónoma Metropolitana. Recuperado de [http://bindani.itzt.uam.mx/concern/parent/zg64tk941/file\\_sets/8623hx730](http://bindani.itzt.uam.mx/concern/parent/zg64tk941/file_sets/8623hx730)
- Lüdecke, D. (2020). *sjPlot: Data Visualization for Statistics in Social Science*. <https://doi.org/10.5281/zenodo.1308157>
- Lüdecke, D. & Ranzolin, D. (2020). sjlabelled: Labelled Data Utility Functions (Versión 1.1.3). Recuperado de <https://CRAN.R-project.org/package=sjlabelled>
- Luján, J. de J. (2009). *Imputación del ingreso en la Encuesta Nacional de Ocupación y Empleo* (Maestría en Ciencias en Estadística Oficial, Centro de Investigación en Matemáticas, A.C). Centro de Investigación en Matemáticas, A.C. Recuperado de <https://cimat.repositorioinstitucional.mx/jspui/bitstream/1008/200/2/TE%20286.pdf>
- Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A. et al. (2020). *robustbase: Basic Robust Statistics*. Recuperado de <http://robustbase.r-forge.r-project.org/>
- Márquez, C., Prieto, V. & Escoto, A. (2020). Segmentación en el ingreso por trabajo según condición migratoria, género y ascendencia étnico-racial en Uruguay. *Migraciones*, (49), 85-118.
- Martínez, J. C. (2017). Muestras complejas con R. Recuperado el 21 de agosto de 2020, de [https://rstudio-pubs-static.s3.amazonaws.com/231846\\_99c9292952fe4e95be25c04bee468e93.html](https://rstudio-pubs-static.s3.amazonaws.com/231846_99c9292952fe4e95be25c04bee468e93.html)
- Medina, F. & Galvan, M. (2008). *Descomposición del coeficiente de Gini por fuentes de ingreso: Evidencia empírica para América Latina 1999-2005*. Santiago Chile: CEPAL.

- Mendenhall, W., Beaver, R. J. & Beaver, B. M. (2014). *Introducción a la probabilidad y estadística* (J. A. Velázquez Arellano, Trad.). México, D.F.: Cengage Learning Editores.
- Meyer, F., Perrier, V. & Caroll, I. (2020). esquisse: Explore and Visualize Your Data Interactively (Versión 0.3.0). Recuperado de <https://CRAN.R-project.org/package=esquisse>
- Milanovic, B. (1997). A simple way to calculate the Gini coefficient, and some implications. *Economics Letters*, 56(1), 45-49. [https://doi.org/10.1016/S0165-1765\(97\)00101-8](https://doi.org/10.1016/S0165-1765(97)00101-8)
- Mincer, J. A. (1974). Introduction to “Schooling, Experience, and Earnings”. *Schooling, Experience, and Earnings*, 1-4.
- Moore, D. S. (2010). *Estadística aplicada básica* (J. Comas, Trad.). Barcelona: Antoni Bosch.
- Negrete, R. (2011). El concepto estadístico de informalidad y su integración bajo el esquema del Grupo de Delhi. *Revista International de Estadística y Geografía*, 2(3), 76.
- Neuwirth, E. (2014). RColorBrewer: ColorBrewer Palettes (Versión 1.1-2). Recuperado de <https://CRAN.R-project.org/package=R-ColorBrewer>
- Ochoa, S. M. (2016). Trayectorias laborales durante la crisis económica 2008-2009 en México. *Economía Informa*, 399, 34-58. <https://doi.org/10.1016/j.ecin.2016.08.004>
- Ospino, C., Roldán, P. & Barraza, N. (2010). La descomposición salarial de Oaxaca-Blinder: Métodos, críticas y aplicaciones. Una revisión de la literatura. *Revista de Economía Del Caribe*, 0(5). <https://doi.org/10.14482/rec.v0i5.1258>
- Pacheco, M. E. (2004). *Ciudad de México, heterogénea y desigual: Un estudio sobre el mercado de trabajo*. Mexico, D.F.: El Colegio de México.
- Pacheco, E. & Parker, S. (2001). Movilidad en el mercado de trabajo urbano: Evidencias longitudinales para dos periodos de crisis en México (Mobility in the Urban Labor Market: Longitudinal Evidence for Two Periods of Crisis in Mexico). *Revista Mexicana de Sociología*, 63(2), 3-26. JSTOR. <https://doi.org/10.2307/3541345>
- Padrón, M., Gandini, L. & Navarrete, E. L. (Eds.). (2017). *No todo el trabajo es empleo: Avances y desafíos en la conceptualización y medición del trabajo en México* (Primera edición). Ciudad de

- México: El Colegio Mexiquense, A.C.; Universidad Nacional Autónoma de México, Instituto de Investigaciones Jurídicas.
- Partida, V. (2011). Estimación indirecta de tasas de ingreso y de retiro de la actividad económica para México. *Estudios Demográficos y Urbanos*, 26(1), 33-73.
- Pedrero, M. (2003). Las condiciones de trabajo en los años noventa en México. Las mujeres y los hombres: ¿ganaron o perdieron? *Revista Mexicana de Sociología*, 65(4), 733-761. JSTOR. <https://doi.org/10.2307/3541581>
- Pedrero, M. (2018). *El trabajo y su medición: Mis tiempos: antología de estudio sobre trabajo y género* (Primera edición). México: Universidad Nacional Autónoma de México, Centro Regional de Investigaciones Multidisciplinarias; Miguel Ángel Porrúa.
- Pedrero, M. & Rendón, T. (1982). El trabajo de la mujer en México en la década de los setenta. En *Estudios sobre la mujer. I. El empleo y la mujer. Bases teóricas, metodológicas y evidencias empíricas*. México, D.F: Secretaría de Programación y Presupuesto.
- R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Recuperado de <https://www.R-project.org/>
- Ram, K., Wickham, H., Richards, C. & Baggett, A. (2018). wesanderson: A Wes Anderson Palette Generator (Versión 0.3.6). Recuperado de <https://CRAN.R-project.org/package=wesanderson>
- Rendón, T. & Salas, C. (1986). La población económicamente activa en el censo de 1980. Comentarios críticos y una propuesta de ajuste. *Estudios Demográficos y Urbanos*, 1(2 (2)), 291-309.
- Rinker, T., Kurkiewicz, D., Hughitt, K., Wang, A., Aden-Buie, G., Wang, A. & Burk, L. (2019). pacman: Package Management Tool (Versión 0.5.1). Recuperado de <https://CRAN.R-project.org/package=pacman>
- Robinson, D. & Hayes, A. (2019). broom: Convert Statistical Analysis Objects into Tidy Tibbles. Recuperado de <https://CRAN.R-project.org/package=broom>
- Rodríguez, E. & López, B. (2015). Imputación de ingresos laborales. Una aplicación con encuestas de empleo en México. *El trimestre económico*, 82(325), 117-146.
- Salas, C. (2003). Trayectorias laborales entre el empleo, el desempleo

- y las microunidades en México. *Papeles de población*, 9(38), 121–157.
- Salas, C. & Zepeda, E. (2003). Empleo y salarios en el México contemporáneo. En E. M. de la Garza & C. Salas (Eds.), *La situación del trabajo en México, 2003* (1. ed, pp. 55-76). México, D.F: Plaza y Valdés.
- Signorelli, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T. et al. (2020). DescTools: Tools for Descriptive Statistics (Versión 0.99.32). Recuperado de <https://CRAN.R-project.org/package=DescTools>
- Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, 103(2684), 677-680.
- Straka, M. & Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88-99. Vancouver, Canada: Association for Computational Linguistics. <https://doi.org/10.18653/v1/K17-3009>
- Wang, Y. (2017). extdplyr: Data Manipulation Extensions of “Dplyr” and “Tidyr” (Versión 0.1.4). Recuperado de <https://CRAN.R-project.org/package=extdplyr>
- Wasserstein, R. L. & Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2), 129-133. <https://doi.org/10.1080/00031305.2016.1154108>
- Wickham, H. & Miller, E. (2019). haven: Import and Export “SPSS”, “Stata” and “SAS” Files (Versión 2.2.0). Recuperado de <https://CRAN.R-project.org/package=haven>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. et al. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- Wikipedia, la enciclopedia libre (2020). Diagrama de caja. Recuperado de [https://es.wikipedia.org/w/index.php?title=Diagrama\\_de\\_caja&oldid=126639841](https://es.wikipedia.org/w/index.php?title=Diagrama_de_caja&oldid=126639841)
- Wooldridge, J. (2010). *Introducción a la Econometría*. 4e. México, México: Cengage Learning Editores. Recuperado de <http://public.eblib.com/choice/publicfullrecord.aspx?p=4641575>

- Xie, Y., Allaire, J. J. & Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Florida, EUA: Chapman and Hall/CRC. Recuperado de <https://bookdown.org/yihui/rmarkdown>
- Yaffee, R. A. (2002). *Robust Regression Analysis: Some Popular Statistical Package Options*.
- Zeileis, A. & Kleiber, C. (2014). *ineq*: Measuring Inequality, Concentration, and Poverty (Versión 0.2-13). Recuperado de <https://CRAN.R-project.org/package=ineq>

# Glosario

**Análisis bivariado:** estudio de la relación que guardan dos variables entre sí. El análisis estadístico cambia de acuerdo con el tipo y escala de medición de las variables.

**Análisis de panel:** tipo de análisis donde se registran múltiples mediciones a lo largo del tiempo para un individuo o unidad de análisis. También se le conoce como análisis prospectivo.

**Análisis de varianza-ANOVA:** modelo estadístico que contrasta si varias poblaciones tienen la misma media, comparando lo separadas que están entre sí las medias muestrales en relación con la variación existente dentro de las muestras (Moore, 2010, p. 661).

**Coeficiente de Gini:** índice que mide la desigualdad, muy utilizado para medirla en distribución de los ingresos. Se calcula comparando la proporción acumulada de la variable población y la proporción acumulada de la variable de riqueza. El coeficiente varía de 0 (o 0%) a 1 (o 100%), donde 0 representa la igualdad perfecta y 1 representa la desigualdad perfecta.

**Correlación (de Pearson):** es una medición de la fuerza y la dirección de la relación lineal entre dos variables cuantitativas. Se abrevia como  $r$  para la muestra y como  $\rho$  para la población. Es la sumatoria del producto de los valores estandarizados entre las dos variables. El coeficiente de Pearson  $r$  puede tener valores entre  $-1$  y  $1$ .

**Curva de Lorenz:** curva que representa la proporción del ingreso total de la población (eje  $y$ ) que el  $x\%$  inferior de la población gana acumulativamente.

**dataframe:** tipo de objeto en  $R$  que permite manipular conjuntos de datos. Son tablas de datos que almacenan información en dos dimensiones. La información, a diferencia de una matriz, puede ser de diferentes tipos (numérica, carácter, lógicos, de factor, entre otros).

**Desviación estándar:** distancia promedio de un valor con respecto a la media de los datos en estudio.

**División sexual del trabajo:** proceso mediante el cual las tareas y responsabilidades se asignan diferenciadamente de acuerdo con el sexo biológico de una persona.

**ENOE:** Encuesta Nacional de Ocupación y Empleo. Es una encuesta trimestral que tiene por objetivo general “obtener información estadística sobre la fuerza de trabajo y las características ocupacionales de la población a nivel nacional, estatal y por ciudades, así como de variables sociodemográficas que permitan profundizar en el análisis de los aspectos laborales” (INEGI, 2020). Del mismo modo se señala que esta fuente posibilita “ampliar la oferta de indicadores de carácter estratégico para el conocimiento cabal de la realidad nacional y la toma de decisiones orientadas a la formulación de políticas laborales” (INEGI, 2019, p. 9).

**Factor de expansión:** es el inverso de la probabilidad de selección de un individuo en una muestra. Es un valor numérico por el cual se debe multiplicar cada uno de los casos de una encuesta para obtener el total de la población.

**Formato ancho:** (en el caso de este texto) formato en el que cada línea es una persona.

**Formato long o largo:** (en el caso de este texto) formato en el que la base construida presenta más de un renglón por persona.

**Gráfica de caja y brazos (o caja y bigotes):** representación gráfica que permite mostrar datos numéricos a través de sus cuartiles. Toma en cuenta el concepto de rango intercuartílico (RIC), que es la diferencia entre el cuartil 1 y el cuartil 3; es decir, establece el rango donde se concentra el 50% de los datos.

**Gráfica de densidad:** representación gráfica donde, en el eje de las  $x$ , normalmente se establecen los valores de la variable cuantitativa  $y$ , y en el eje de las  $y$ , la densidad de probabilidad. En los paquetes estadísticos la densidad se construye a partir de modelaciones de la frecuencia relativa.

**Gráfico de tallo y hoja:** representación gráfica que consiste en un “tallos” establecido por decenas, o algún múltiplo de ellas, y “hojas” que representan cada uno de los datos en análisis que permiten observar la distribución de la variable en estudio.

**Gráfica aluvial:** representación gráfica también conocida como gráfica de hilos. Es un tipo de Diagrama de Sankey que sirve para

representar los flujos de datos y su magnitud. **Heterocedasticidad:** condición donde un conjunto de datos no tiene una varianza común y constante. En específico, en el caso de los modelos estadísticos, se refiere a que los errores no tienen una varianza constante y común. Como resultado, los errores estándar de los coeficientes se subestiman.

**Histograma:** representación gráfica de barras que representan intervalos de una variable cuantitativa y sus frecuencias, frecuencias relativas o densidad. Las barras no están separadas puesto que representan intervalos que se unen en sus límites y, por lo general, son del mismo color.

**Hogar:** conjunto de personas que residen habitualmente en la misma vivienda particular y se sostienen de un gasto común, principalmente para alimentación. “Una organización estructurada a partir de lazos sociales entre personas unidas o no por relaciones de parentesco que comparten una misma vivienda (una sola persona también puede formar un hogar” (INEGI, 2019, p. 21).

**Imputación:** cálculo de un valor a partir de un modelo o condiciones lógicas.

**Inferencia:** proceso de estimación de parámetros poblacionales con información de la muestra aleatoria.

**Ingresos por trabajo:** percepción monetaria (salarios, honorarios, pagos por comisiones, entre otros) que la población ocupada obtiene o recibe del trabajo que desempeñó en la semana de referencia. Los ingresos se calculan en forma mensual.

**Media aritmética:** es la suma de todos los valores de nuestra variable dividida entre el total de observaciones. La media tiene varias propiedades; por ejemplo, si sumamos todas las desviaciones a este valor, la suma de ellas es cero.

**Mediana:** medida de posición y tendencia central que presenta el valor de la variable en la posición central en un conjunto de datos ordenados. Es decir, supera al 50% de los casos y su valor es superado por el 50% restante.

**Microdatos:** es la información recolectada en las encuestas en la unidad más pequeña de registro. Por ejemplo, la información a nivel de las personas de la ENOE se considera microdatos, así como la del nivel de hogar y la vivienda.

**Moda:** es el valor de una variable que más se repite dentro de un conjunto de observaciones.

**Ocupación principal:** es la ocupación que el entrevistado o entrevistada en la ENOE reporta como principal, ya sea en tiempo de trabajo o en términos de ingresos.

**outliers u observaciones atípicas:** son los valores individuales que quedan fuera del aspecto general de la distribución de una variable.

**PEA:** Población Económicamente Activa. Se refiere a la población en edad de trabajar que está ocupada, o bien, que está buscando trabajo activamente en un periodo estipulado antes de la entrevista.

**PET:** Población en Edad de Trabajar. Se refiere a la población que tiene 15 años y más, para el caso específico de México. Este límite etario está legislado por cada país.

**PNEA:** Población No Económicamente Activa clasificada en esta categoría [PNEA]. Población en edad de trabajar que no está ocupada en el mercado de trabajo y que además no intenta modificar esa condición de no ocupación involucrándose en el mercado a través de la búsqueda activa. Dentro de este grupo se puede identificar su condición de disponibilidad para trabajar, por lo que también se diferencia entre PNEA disponible y PNEA no disponible. También suele estudiarse según la identificación del tipo de actividades no económicas que realizan (si son estudiantes, se dedican a quehaceres domésticos, o son pensionados y/o jubilados, etcétera).

**Prueba de independencia- chi-cuadrado de Pearson:** es el cálculo que permite comprobar la independencia de frecuencias entre dos variables aleatorias. Es decir, permite observar el grado de relación que tienen dos variables entre sí o si éstas son completamente independientes.

**Prueba Kruskal-Wallis:** prueba no paramétrica, similar a la ANOVA, donde los cálculos se realizan sobre las medidas de rangos de las variables y no los valores originales.

**Regresión lineal:** es un modelo estadístico que permite observar la relación entre dos o más variables, asumiendo linealidad entre la variable dependiente y las variables independientes.

**Segregación:** distancias (espaciales o sociales) entre dos grupos por su adscripción; en este caso, por la condición de ser hombre o mujer (Guzmán, 2009; Kuri, 2014).

**Tablas de doble entrada:** también llamadas tablas de contingencia.

Tablas en las que las columnas almacenan los valores de una variable categórica y en las filas se almacenan los valores de una segunda variable. En las celdas se establecen los conteos de todas las combinaciones posibles entre los valores de una variable con la otra. La suma de los marginales de las columnas es igual a la suma de los marginales de las filas y este valor es el total de observaciones.

**Tasa de participación económica:** cociente entre la PEA y la PET.

**Tasa de transición:** medición de la probabilidad de cambiar de un estado  $i$  al estado  $j$ . Se presenta como una manera de analizar la información longitudinal.

**Variable cualitativa:** tipo de variable que describe las cualidades, circunstancias o características de un objeto o persona. Estos atributos expresados son categorías no numéricas. Una variable cualitativa puede ser nominal, ordinaria o binaria.

**Variable cuantitativa:** un tipo de variable estadística que otorga un valor numérico a las mediciones registradas. Las variables cuantitativas pueden ser discretas o continuas.

**Varianza:** es una medida de distancia al cuadrado promedio que permite observar cómo varía un conjunto de datos respecto de su media aritmética.

# Anexo: Análisis de texto de las investigaciones en México

## Bibliografía analizada

- Aguilera, A. (2016). El mercado de trabajo y la desigualdad salarial en México y sus regiones, 1992-2014. En L. Estrada Quiroz (Ed.), *Problemas teórico metodológicos en la investigación. Tesis doctorales en Ciencias Sociales* (pp. 153-163). Recuperado de [https://www.academia.edu/28676633/Problemas\\_teo\\_rico\\_metodolo\\_gicos\\_en\\_la\\_Investigaci%C3%B3n\\_Tesis\\_doctorales\\_en\\_Ciencias\\_Sociales](https://www.academia.edu/28676633/Problemas_teo_rico_metodolo_gicos_en_la_Investigaci%C3%B3n_Tesis_doctorales_en_Ciencias_Sociales)
- Aguilera, A. & Castro, D. (2018a). Calificación laboral y desigualdad salarial: Un ejercicio metodológico por conglomerados. *Economía: teoría y práctica*, (49), 65-91. <https://doi.org/10.24275/etypuam/ne/492018/aguilera>
- Aguilera, A. & Castro, D. (2018b). NAFTA and Wage Inequality in Mexico: An Analysis for Border Cities, 1992-2016. *Frontera Norte*, 30(60), 85–110.
- Alcaraz, C., Chiquiar, D. & Salcedo, A. (2012). Remittances, schooling, and child labor in Mexico. *Journal of Development Economics*, 97(1), 156-165. <https://doi.org/10.1016/j.jdeveco.2010.11.004>
- Alcaraz, C., Chiquiar, D. & Salcedo, A. (2015). *Informality and segmentation in the Mexican labor market* (Working Papers, No. 2015-25). Banco de México. Recuperado de <https://www.econstor.eu/bitstream/10419/129955/1/84351468X.pdf>
- Altamirano, J. A., Gutiérrez, L. & Castro, D. (2014). Una perspectiva de la pobreza por ingresos de los trabajadores ocupados en el sector terciario: El caso de doce ciudades mexicanas, 2000-2010. En D. Castro & R. E. Rodríguez (Coords.), *El mercado laboral frente a las transformaciones económicas en México* (pp. 71-90). México: Universidad Autónoma de Coahuila; Plaza y Valdés.

- Recuperado de <https://www.cise.uadec.mx/downloads/LibrosElectrónicos/LibroMercadoLaboral2014.pdf>
- Andrés, R., Czarnecki, L. & Mendoza, M. Á. (2019). A spatial analysis of precariousness and the gender wage gap in Mexico, 2005-2018. *The Journal of Chinese Sociology*, 6(1), 13. <https://doi.org/10.1186/s40711-019-0104-2>
- Ariza, J. & Raymond, J. L. (2018). Technical change and employment in Brazil, Colombia, and Mexico. Who are the most affected workers? *International Labour Review*, 159(2), 137-159. <https://doi.org/10.1111/ilr.12104>
- Ariza, M. & Oliveira, O. de (2014). Terciarización de la mano de obra y protección laboral de la población asalariada en México, 2013. *Realidad, datos y espacio. Revista internacional de estadística y geografía*, 5(2), 34-47.
- Ariza, M. & Oliveira, O. de. (2013). Viejos y nuevos rostros de la precariedad en el sector terciario, 1995-2010. En C. Rabell, *Los mexicanos: Un balance del cambio demográfico*. Ciudad de México: Fondo de Cultura Económica.
- Ayala, E. A. & Chapa, J. C. (2019). Demanda agregada y desigualdad regional por género en México. *Cuadernos de Economía*, 38(77), 399-24. <https://doi.org/10.15446/cuad.econ.v38n77.66561>
- Badillo, G. (2018). Condiciones laborales y actividades económicas en la vejez: Un análisis a partir de la Encuesta Nacional de Ocupación y Empleo. *Coyuntura Demográfica*, 13, 68-77.
- Bargain, O. & Kwenda, P. (2010). *Is Informality Bad? Evidence from Brazil, Mexico and South Africa* (SSRN Scholarly Paper Núm. ID 1545138). Recuperado de Social Science Research Network website: <https://papers.ssrn.com/abstract=1545138>
- Barrandey, J. A. & Barajas, H. A. (2018). Análisis de la productividad en la industria manufacturera e informalidad laboral en la frontera norte de México. En D. Castro & R. E. Rodríguez (Coords.), *Mercado laboral: México y frontera norte*. México: Universidad Autónoma de Coahuila; Ediciones de Laurel. Recuperado de <https://www.cise.uadec.mx/downloads/LibrosElectrónicos/LibroMercadoLaboral2018.pdf>
- Barrón, M. A. (2018). La brecha laboral rural en México. Una grieta invisible de la desocupación. *Economía UNAM*, 15(45), 89-107. Re-

- cuperado de <http://www.scielo.org.mx/pdf/eunam/v15n45/1665-952X-eunam-15-45-89.pdf>
- Benita, F. (2014). A Cohort Analysis of the College Premium in Mexico. *Latin American Journal of Economics*, 51(1), 147-178. Recuperado de [https://www.researchgate.net/publication/263466302\\_A\\_Cohort\\_Analysis\\_of\\_the\\_College\\_Premium\\_in\\_Mexico](https://www.researchgate.net/publication/263466302_A_Cohort_Analysis_of_the_College_Premium_in_Mexico)
- BenYishay, A. & Pearlman, S. (2013). *Homicide and Work: The Impact of Mexico's Drug War on Labor Market Participation* (SSRN Scholarly Paper Núm. ID 2302437). <https://doi.org/10.2139/ssrn.2302437>
- Bernal, R., Valenzuela, J. A. & Lara, B. E. (2016). Desocupación en la frontera norte de México. Consecuencias en las personas mayores de cuarenta años. *Estudios Sociales. Revista de Alimentación Contemporánea y Desarrollo Regional*, 26(48), 305-332.
- Bertoli, S. & Murard, E. (2020). Migration and co-residence choices: Evidence from Mexico. *Journal of Development Economics*, 142, 102330. <https://doi.org/10.1016/j.jdeveco.2019.01.011>
- Bobba, M., Flabbi, L. & Levy, S. (2018). *Labor Market Search, Informality and Schooling Investments*. Recuperado de [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3305507](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3305507)
- Bonilla, R. (2015). Informalidad y precariedad laboral en el Distrito Federal. La economía de sobrevivencia. *Economía Informa*, 351, 69-84.
- Botello, J. (2011). Algunos indicadores del mercado de trabajo. *Análisis Económico*, XXVI(63), 247-263.
- Bouchot, J. A. (2018). *The unintended distributional consequences of the 2012 rise in the Mexican minimum wage*. Recuperado de <https://www.aiel.it/cms/cms-files/submission/all20170615235404.pdf>
- Brussolo, M. E. (2017). Effects on Inequality of a Radical Wave of Reforms: The Mexican Case. *Revista Internacional de Ciencias Sociales y Humanidades*, XXVII(2), 27-48. Recuperado de <https://www.redalyc.org/pdf/654/65456039003.pdf>
- Cacciamali, M. C. & Tatei, F. (2013). Género y salarios de la fuerza de trabajo calificada en Brasil y México. *Problemas del desarrollo*. Recuperado de [http://www.scielo.org.mx/scielo.php?pid=S0301-70362013000100004&script=sci\\_abstract&tlang=fr](http://www.scielo.org.mx/scielo.php?pid=S0301-70362013000100004&script=sci_abstract&tlang=fr)

- Calderón, C., Huesca, L. & Ochoa, G. L. (2017). Análisis comparativo de la desigualdad salarial entre México y Estados Unidos. *Investigación Económica*, 76(300), 3-31.
- Calderón, C., Ochoa, G. L. & Huesca, L. (2017). Mercado laboral y cambio tecnológico en el sector manufacturero mexicano (2005-2014). *Economía, sociedad y territorio*, XVII(54), 523-560.
- Camberos, M. & Castro, A. E. (2018). Desempleo, salarios e informalidad: Un análisis de las entidades de México y la frontera norte. En D. Castro & R. E. Rodríguez (Coords.), *Mercado laboral: México y frontera norte*. México: Universidad Autónoma de Coahuila; Ediciones de Laurel. Recuperado de <https://www.cise.uadec.mx/downloads/LibrosElectonicos/LibroMercadoLaboral2018.pdf>
- Campos, R. M. (2010). The effects of macroeconomic shocks on employment: The case of Mexico. *Estudios Económicos*, 25(1), 177-246. México: El Colegio de México.
- Campos, R. M. (2015). El salario mínimo y el empleo: Evidencia internacional y posibles impactos para el caso mexicano. *Economía UNAM*, I2(36), 90-106.
- Campos, R. M. & Lustig, N. (2017). *Labour income inequality in Mexico: Puzzles solved and unsolved*, WIDER Working Paper, No. 2017/186. Recuperado de <https://www.econstor.eu/bitstream/10419/190031/1/wp2017-186.pdf>
- Campos, R. M. & Rodríguez, J. A. (2011). Trade and Occupational Employment in Mexico since NAFTA. *OECD Trade Policy Papers*, (129). <https://doi.org/10.1787/5kg3nh5q7p5k-en>
- Campos, R. M., Esquivel, G. & Santillán, A. S. (2017). El impacto del salario mínimo en los ingresos y el empleo en México. *Revisada de la CEPAL*, (122), 205-234.
- Campos, R. M., Hincapié, A. & Rojas, R. I. (2012). Family income inequality and the role of married females' earnings in Mexico: 1988-2010. *Latin American Journal of Economics*, 49(1), 67-98.
- Canales, R. A., Román, Y. G. & Ovando, W. (2017). Emprendimiento de la población joven en México. Una perspectiva crítica. *Entreciencias: Diálogos en la Sociedad del Conocimiento*, 5(12). <http://dx.doi.org/10.21933/J.EDSC.2017.12.211>
- Cano, J. (2015). The role of the informal sector in the early careers of

- less-educated workers. *Journal of Development Economics*, 112, 33-55. <https://doi.org/10.1016/j.jdeveco.2014.10.002>
- Cano, J. (2016). Informal Labor Markets and On-the-job Training: Evidence from Wage Data. *Economic Inquiry*, 54(1), 25-43. <https://doi.org/10.1111/ecin.12279>
- Cantú, R., Gómez, A. S. & Villarreal, H. J. (2016). The Labor-Market Deterioration and its Relation with Poverty during the International Crises in Mexico. *Realidad, datos y espacio revista internacional de estadística y geografía*, 7(3), 25-29.
- Cardero, M. E. & Mendoza, M. A. (2012). Women's Industrial Employment in Mexico, Measures of Segregation and Discrimination. *Journal of Business and Economics*, 3(6), 410-423.
- Cardero, M. E., Mendoza, M. Á. & Galán, P. (2015). The employment of women in the manufacturing industry after NAFTA. Discrimination and segregation. *Global Journal of Human-Social Science: A Arts & Humanities - Psychology*, 15(2). Recuperado de [https://www.researchgate.net/profile/Miguel\\_Mendoza\\_Gonzalez2/publication/279725959\\_The\\_Employment\\_of\\_Women\\_in\\_the\\_Manufacturing\\_Industry\\_After\\_NAFTA\\_Discrimination\\_and\\_Segregation/links/5598a87b08ae5d8f3933f96f/The-Employment-of-Women-in-the-Manufacturing-Industry-After-NAFTA-Discrimination-and-Segregation.pdf](https://www.researchgate.net/profile/Miguel_Mendoza_Gonzalez2/publication/279725959_The_Employment_of_Women_in_the_Manufacturing_Industry_After_NAFTA_Discrimination_and_Segregation/links/5598a87b08ae5d8f3933f96f/The-Employment-of-Women-in-the-Manufacturing-Industry-After-NAFTA-Discrimination-and-Segregation.pdf)
- Carrillo, S. & Ríos, J. G. (2014). Oferta de trabajo de los estudiantes de la Universidad de Guadalajara y de México: Un análisis comparativo. *Perfiles educativos*, 36(144), 85-104.
- Casanueva, C. (2013). Mexico's Universal Telecommunications Service Policies and Regulatory Environment in an International Perspective, 1990-2010. *Journal of Information Policy*, 3, 267-303.
- Castillo, D. & Vela, F. (2013). Movilidad laboral y transmisión intergeneracional del autoempleo informal en México. *Gaceta laboral*, 19(1), 5-35.
- Castro, D. & Rodríguez, R. E. (2014). Dispersión salarial en la industria automotriz: Un estudio comparativo para las ciudades de Saltillo y Hermosillo. En D. Castro y R. Rodríguez (Coords.), *El mercado laboral frente a las transformaciones económicas en México*, (pp. 257-286). México: Universidad Autónoma de Coahuila; Plaza

- y Valdés. Recuperado de <https://www.cise.uadec.mx/downloads/LibrosElectonicos/LibroMercadoLaboral2014.pdf>
- Castro, D., Rodríguez, R. E. & Brown, F. (2018). La brecha salarial por género y recesión económica en la frontera norte de México. En D. Castro & R. E. Rodríguez (Coords.), *Mercado laboral: México y frontera norte*. México: Universidad Autónoma de Coahuila; Ediciones de Laurel. Recuperado de <https://www.cise.uadec.mx/downloads/LibrosElectonicos/LibroMercadoLaboral2018.pdf>
- Castro, D., Rodríguez, R. E. & Huesca, L. (2013). La calificación laboral en ocupaciones tecnológicas y no tecnológicas en México y sus regiones. *Estudios Sociales. Revista de Alimentación Contemporánea y Desarrollo Regional*, 21(42), 87-112.
- Castro, N., Escoto, A. R. & Pacheco, E. (2017). Transformaciones en la medición del “trabajo en la ocupación”. Una revisión de la XIX CIET. En M. Padrón, L. Gandini & E. L. Navarrete, *No todo el trabajo es empleo. Avances y desafíos en la conceptualización y medición del trabajo en México*. El Colegio Mexiquense; Instituto de Investigaciones Jurídicas-UNAM.
- Conover, E., Khamis, M. & Pearlman, S. (2018). *Reversed Migration Trends and Local Labor Markets*. Presentado en 2nd IZA/World Bank/NJD Conference on Jobs and Development: Improving Jobs Outcomes in Developing Countries, Washigton, DC. Recuperado de [http://conference.iza.org/conference\\_files/worldbank\\_2019/khamis\\_m3385.pdf](http://conference.iza.org/conference_files/worldbank_2019/khamis_m3385.pdf)
- Contreras, A. B. & Rubio, J. A. (2014). Las tendencias del capital humano en México: Un análisis de la trayectoria de la educación por género 2005-2012. *Entreciencias: Diálogos en la Sociedad del Conocimiento*, 2(3), 67-80. <https://doi.org/10.21933/J.EDSC.2014.03.032>
- Coronado, R. & Saucedo, E. (2019). Drug-related violence in Mexico and its effects on employment. *Empirical Economics*, 57(2), 653-681. <https://doi.org/10.1007/s00181-018-1458-z>
- Cortes, G. M. & Morris, D. M. (2020). *Are routine jobs moving south? Evidence from changes in the occupational structure of employment in the USA and Mexico* (2a ed.). WIDER Working Paper 2020/11. <https://doi.org/10.35188/UNU-WIDER/2020/768-2>
- Cota, R. & Navarro, A. (2015). Análisis del mercado laboral y el empleo informal mexicano. *Papeles de población*, 21(85), 211-

249. Recuperado de [http://www.scielo.org.mx/scielo.php?script=sci\\_arttext&pid=S1405-74252015000300008](http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-74252015000300008)
- Cunningham, W. V. (2001). *Breadwinner or caregiver? How household role affects labor choices in Mexico*. Policy Research Working Paper, No. 2743, World Bank. Recuperado de [https://www.researchgate.net/publication/23549404\\_Breadwinner\\_or\\_Caregiver\\_How\\_Household\\_Role\\_Affects\\_Labor\\_Choices\\_in\\_Mexico](https://www.researchgate.net/publication/23549404_Breadwinner_or_Caregiver_How_Household_Role_Affects_Labor_Choices_in_Mexico)
- De Hoyos, R., Gutiérrez, C. & Vargas, J. V. (2016). *Idle youth in Mexico: Trapped between the war on drugs and economic crisis*, Policy Research Working Paper, No. 7558, World Bank. Recuperado de World Bank Group website: <https://documents.worldbank.org/en/publication/documents-reports/documentdetail>
- Díaz, J. C., Guerra, A. & Cruz, N. (2016). Un modelo de red bayesiana de la informalidad laboral en Veracruz orientado a una simulación social basada en agentes. *Research in Computing Science*, 113, 157-170.
- Dorn, F. & Silbersdorff, A. (2017). The Impact of Unpaid Work on Employment Status in Mexico. *CEGE Discussion Papers*, (328). <https://doi.org/10.2139/ssrn.3080932>
- Dougherty, S. & Escobar, O. (2013). The Determinants of Informality in Mexico's States. *OECD Economics Department Working Papers*, (1043). Recuperado de <https://www.oecd-ilibrary.org/docserver/5k483jrvnjq2-en.pdf?expires=1565648906&id=id&accname=guest&checksum=EEED1DA2500D0209E86101883C-714DAE>
- Duval, R. & Orraca, P. (2009). A Cohort Analysis of Labor Participation in Mexico, 1987-2009. *IZA Discussion Paper*, (4371). Recuperado de <http://ftp.iza.org/dp4371.pdf>
- Duval, R. & Smith, R. (2011). *Informality and Seguro Popular under Segmented Labor Markets*. Recuperado de [http://conference.iza.org/conference\\_files/worldb2011/3229.pdf](http://conference.iza.org/conference_files/worldb2011/3229.pdf)
- Duval, R., Fields, G. S. & Jakubson, G. H. (2015). *Analysing income distribution changes: Anonymous versus panel income approaches* (26a ed.). WIDER Working Paper 2015/026. <https://doi.org/10.35188/UNU-WIDER/2015/911-4>
- Escoto, A. R. & García, B. (2015). Condiciones laborales y comercio exterior en México. En D. Castillo, N. Baca & R. Todaro

- (Coords.), *Trabajo global y desigualdades en el mercado laboral* (pp. 91-134). México: Universidad Autónoma del Estado de México; Facultad de Ciencias Políticas y Sociales. Recuperado de <http://ri.uaemex.mx/bitstream/handle/20.500.11799/66128/TrabajoGlobal.pdf?sequence=1>
- Escoto, A. R., Márquez, C. & Prieto, V. (2017). Desempleo abierto y desalentado en tres mercados de trabajo latinoamericanos. En S. M. Ochoa & R. P. Román (Coords.), *Población y mercados de trabajo en América Latina. Temas emergentes* (pp. 81-11). México: Instituto de Investigaciones Jurídicas-UNAM. Recuperado de <http://ru.juridicas.unam.mx/xmlui/handle/123456789/13168>
- Fareed, F., Gabriel, M., Lenain, P. & Reynaud, J. (2017). Financial Inclusion and Women Entrepreneurship: Evidence from Mexico. *OECD Economics Department Working Papers*, No. 1411). <https://doi.org/10.1787/2fbd0f35-en>
- Félix, G. & Torres, A. J. (2018). Prima salarial al uso de la computadora en el trabajo. Evidencia de microdatos para México. *El trimestre económico*, 85(337), 137-168.
- Flores, I. M. (2012). *The informal sector in Mexico: Implications on health insurance coverage and education* (Doctor of Philosophy, New York University). Recuperado de <https://search-proquest-com.pbidi.unam.mx:2443/pqdtglobal/docview/1264397929/full-textPDF/6EE1B2C96DA249CCPQ/1?accountid=14598>
- Flores, J. G. R., Zamora, S. & Contreras, E. (2013). Transiciones entre el trabajo formal e informal y medios de intermediación laboral en México 2005-2010. *Banco Interamericano de Desarrollo*. Recuperado de [https://pdfs.semanticscholar.org/8287/580e2856acac52dca91083fc355159cd-d2c90.pdf?\\_ga=2.148436671.925693671.1573865334-732215748.1573865334](https://pdfs.semanticscholar.org/8287/580e2856acac52dca91083fc355159cd-d2c90.pdf?_ga=2.148436671.925693671.1573865334-732215748.1573865334)
- Florez, N. & Pacheco, E. (2017). Entre la invisibilización del trabajo de autoconsumo de bienes y la visibilización del trabajo no remunerado. En M. Padrón, L. Gandini & E. L. Navarrete, *No todo el trabajo es empleo. Avances y desafíos en la conceptualización y medición del trabajo en México*. México: El Colegio Mexiquense; Instituto de Investigaciones Jurídicas-UNAM.
- Freije, S., López, G. & Rodríguez, E. (2011). *Effects of the 2008-09*

- economic crisis on labor markets in Mexico.* Policy Research Working Paper, WPS 5840. World Bank. Recuperado de The World Bank website: <http://documents.worldbank.org/curated/en/752221468279886955/Effects-of-the-2008-09-economic-crisis-on-labor-markets-in-Mexico>
- Galhardi, R. (2012). La situación del empleo en la crisis en México. En E. de la Garza (Coord.), *La situación del trabajo en México, 2012, el trabajo en la crisis* (pp. 65-90). México: Plaza y Valdés. Recuperado de <http://www2.itz.uam.mx/sotraem/NovedadesEditoriales/Situaciondeltrabajo.pdf>
- Gallegos, A. (2015). Trabajo infantil en México. Perfil sociodemográfico de los niños trabajadores de 5-11 años de edad. En D. Castillo, N. Baca & R. Todaro (Coords.) *Trabajo global y desigualdades en el mercado laboral* (pp. 277-303). México: Universidad Autónoma del Estado de México; Facultad de Ciencias Políticas y Sociales. Recuperado de <http://ri.uaemex.mx/bitstream/handle/20.500.11799/66128/TrabajoGlobal.pdf?sequence=1>
- Gámez, J. & Rosas, E. (2015). Determinantes de la diferenciación salarial en México. *Multidisciplina. Revista electrónica de la Facultad de Estudios Superiores Acatlán*, (20), 53-75. Recuperado de <http://revistas.unam.mx/index.php/multidisciplina/article/view/55046>
- Garay, S. (2012). Diferencias estatales y regionales en el empleo rural femenino en México. *Estudios Demográficos y Urbanos*, 27(3), 621-659.
- García, S. G. (2017). *Globalization, location and localization of manufacturing employment, and urban wages in Mexico* (thesis). The University of Texas at Austin. <https://doi.org/10.15781/T29C6SH1N>
- García, B. (2012). La precarización laboral y el desempleo en México (2000-2009). En E. de la Garza (Coord.), *La situación del trabajo en México, 2012, el trabajo en la crisis* (pp. 91-118). México: Plaza y Valdés. Recuperado de <http://www2.itz.uam.mx/sotraem/NovedadesEditoriales/Situaciondeltrabajo.pdf>
- García, B. (2010). Inestabilidad laboral en México: El caso de los contratos de trabajo. *Estudios Demográficos y Urbanos*, 25(1), 73-101.

- García, G. (2017). *Desigualdad y polarización salarial en el empleo formal e informal: El caso de México, 2005-2014.* (Doctorado en Ciencias Económicas), Universidad Autónoma Metropolitana-Iztapalapa. Recuperado de <http://148.206.53.84/tesiuami/UAMI22475.pdf>
- García, R. E. (2014). *La llegada del primer hijo: Cambios en el uso del tiempo de los miembros de la pareja en México 2010-2013. Un análisis con la ENOE* (tesis de maestría). El Colegio de México. Recuperado de [https://colmex.userservices.exlibrisgroup.com/view/delivery/52COLMEX\\_INST/1264963100002716](https://colmex.userservices.exlibrisgroup.com/view/delivery/52COLMEX_INST/1264963100002716)
- García, R. & Prudencio, D. (2017). A long-term employment derivation index for Mexico. *Estudios Económicos*, 32(1), 133-165.
- Gaxiola, S. C., Márquez, C. & Montoya, M. V. J. (2017). Del desempleo a la desocupación. Alcances y limitaciones de los cambios en la medición de la fuerza de trabajo sobre las estadísticas de desocupación en México. En M. Padrón, L. Gandini & E. L. Navarrete, *No todo el trabajo es empleo. Avances y desafíos en la conceptualización y medición del trabajo en México*. México: El Colegio Mexiquense; Instituto de Investigaciones Jurídicas- UNAM.
- González, A. & Uribe, J. A. (2018). Precarización del empleo en Morelia., Michoacán, México. *CIMEXUS*, 13(1), 31-50.
- González, G. (2017). *Mercado de trabajo en México. Acumulación, salario y ganancia.* (Primera edición). Ciudad de México: Universidad Nacional Autónoma de México.
- Guerra y Guerra, G. G. (2012). Políticas de protección laboral en etapas de crisis en México: El Seguro de Desempleo del Distrito Federal. *Estudios Demográficos y Urbanos*, 27(3), 661-698.
- Guerrero, C. M., Molina, M. & Colchero, M. A. (2017). Employment changes associated with the introduction of taxes on sugar-sweetened beverages and nonessential energy-dense food in Mexico. *Preventive medicine*. Recuperado de <https://www.sciencedirect.com/science/article/pii/S0091743517303249>
- Guillermo, S. B. & Harberger, A. C. (2012). Measuring the social opportunity cost of labor in Mexico. *Journal of Benefit-Cost Analysis*, 3(2). Recuperado de <https://www.cambridge.org/core/journals/journal-of-benefit-cost-analysis/article/measuring-the-social-opportunity-cost-of-labor-in-mexico/32CE-310518D5581EF0EA8A92CC950589>

- Hernández, J. S., Desidério, E. D. J. & & Aguillar, N. (2019). Exploratory Study on the Determinants of Informal Employment in the Current Mexican Return Migration. *American International Journal of Social Science*, 8(1). <https://doi.org/10.30845/aijss.v8n1p10>
- Hernández, A. K. & Vargas, E. D. (2016). Condiciones del trabajo estudiantil urbano y abandono escolar en el nivel medio superior en México. *Estudios Demográficos y Urbanos*, 31(3), 663-696.
- Horbath, J. E. & Gracia, A. (2014). Discriminación laboral y vulnerabilidad de las mujeres frente a la crisis mundial en México. *Economía, sociedad y territorio*, 14(45), 465-495.
- Huesca, L. (2010). El empleo informal en la frontera norte de México y el caso de Sonora: Un análisis de expectativas en los ingresos. *Región y Sociedad*, 22(49), 1743.
- Huesca, L. & Gonzalez, J. (2017). Technical Change and Labor Market in Mexico: The Shape of Occupation and Qualification for a Better Assessment. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2943160>
- Huesca, L. & Ochoa, G. (2016). Desigualdad salarial y cambio tecnológico en la frontera norte de México. *Problemas del Desarrollo*, 47(187), 165-188. <https://doi.org/10.1016/j.rpd.2016.10.006>
- Ibarra, C. A. & Ros, J. (2017). *The decline of the labour share in Mexico, 1990-2015* (183a ed.). WIDER Working Paper 2017/183. <https://doi.org/10.35188/UNU-WIDER/2017/409-4>
- Baron, J. D., Popova, A. & Sánchez, A. (2016). *Following Mexican Youth: A Short-Run Study of Time Use Decisions*. <https://doi.org/10.1596/1813-9450-7534>
- Kuri, I. (2014). Mujeres y mercados de trabajo: Análisis de la segregación ocupacional por sexo en México. *International Journal of Innovation and Applied Studies*, 9(1), 279-286. Recuperado de <https://pdfs.semanticscholar.org/48c7/8cfb0bb4de3c2a94ebc-0311607082deba097.pdf>
- Liquitaya, J. D. & Gutiérrez, G. (2011). Fluctuaciones del producto y variaciones asimétricas de la ocupación en México: 2000:2-2009:4. *Denarius*, 23(47). Recuperado de <https://denarius.itz.uam.mx/index.php/denarius/article/view/119>

- Llamas, I. & Hernández, J. M. (2014). Mercado laboral y determinantes del salario en los trabajadores de la educación: México 1995-2010. En D. Castro & R. E. Rodríguez (Coords.), *El mercado laboral frente a las transformaciones económicas en México* (pp. 135-172). México: Universidad Autónoma de Coahuila; Plaza y Valdés. Recuperado de <https://www.cise.uadec.mx/downloads/LibrosElectonicos/LibroMercadoLaboral2014.pdf>
- López, J. & Peláez, Ó. (2015). El desigual impacto de la crisis económica de 2008-2009 en los mercados de trabajo de las regiones de México: La frontera norte frente a la región sur. *Contaduría y Administración*, 60, 195-218. <https://doi.org/10.1016/j.cya.2015.05.004>
- López, L. F. & Levy, S. (2016). Labor Earnings, Misallocation, and the Returns to Education in Mexico. *IDB Publications (Working Papers)* 7454. Recuperado de Inter-American Development Bank website: <https://ideas.repec.org/p/ida/brikps/7454.html>
- Luyando, J. R. (2017). Condiciones laborales de niños y jóvenes asalariados en México 2006-2014. *Revista de Ciencias Sociales*, VI(154), 47-62. <https://doi.org/10.15517/rcc.v0i154.29193>
- Luyando, J. R. (2018). The Working Conditions of Children Assumed to Perceive an Income in Mexico: A Comparison between 2007 and 2013. *International Journal of Humanities and Social Science*, 8(7), 180-189.
- Méndez, A. V., Sánchez, E. E. & Castro, D. (2018). Efectividad de los mecanismos de búsqueda de empleo en el mercado laboral mexicano. *Ensayos de Economía*, 28(52), 77-100. <https://doi.org/10.15446/ede.v28n52.72369>
- Mendoza, J. E. (2014). The impact of return migration on the Mexican labor market. *RIEM. Revista internacional de estudios migratorios*, 4(2), 183-205. Recuperado de <http://ojs.ual.es/ojs/index.php/RIEM/article/view/401>
- Mendoza, M. Á., Cardero, M. E. & Ortiz, A. S. (2017). Algunos hechos estilizados y explicativos sobre el diferencial y la discriminación salarial por sexo en México, 1987-2015. *Investigación Económica*, 76(301), 103-135.

- Mendoza, J. E. (2012). Características y determinantes de los cambios recientes de los flujos migratorios de trabajadores mexicanos hacia Estados Unidos. *Papeles de población*. Recuperado de [http://www.scielo.org.mx/scielo.php?pid=S1405-74252012000100003&script=sci\\_arttext&tlang=pt](http://www.scielo.org.mx/scielo.php?pid=S1405-74252012000100003&script=sci_arttext&tlang=pt)
- Román, Y. G., Montoya, B. J., Lozano, D. & Gaxiola, S. C. (2019). Mortalidad según tipos de ocupación en México, 2014. *Población y Salud en Mesoamérica, 17(1)*. Recuperado de <https://dialnet.unirioja.es/servlet/articulo?codigo=7015207>
- Montoya, M. V. J. (2019). Cambio demográfico y proveeduría laboral de los hogares en las urbes de México, 2005-2017. *Revista Latinoamericana de Población, 13(24)*, 63-81. <https://doi.org/10.31406/relap2019.v13.i1.n24.3>
- Mora, M. & De Oliveira, O. (2012). Los dilemas de la integración laboral juvenil en tiempos de crisis. En E. de la Garza (Coord.), *La situación del trabajo en México, 2012, el trabajo en la crisis* (pp. 155-191). México: Plaza y Valdés. Recuperado de <http://www2.izt.uam.mx/sotraem/NovedadesEditoriales/Situaciondeltrabajo.pdf>
- Mungaray, A., Osuna, J. G., Ramírez, M., Ramírez, N. & Escamilla, A. (2015). Emprendimientos de micro y pequeñas empresas mexicanas en un escenario local de crisis económica: El caso de Baja California, 2008-2011. *Frontera norte, 27(53)*, 115-146.
- Murrieta, P. (2016). Child labor and household composition: Determinants of child labor in Mexico. *Asian Journal of Latin American Studies, 29(3)*, 29-54. Recuperado de <https://bettercarenetwork.org/sites/default/files/ChildLab.pdf>
- Navarrete, E. L. & Caro, N. R. (2014). La vulnerabilidad de las mujeres en el trabajo y la salud en México. En E. L. Navarrete & N. R. Caro (Coords.), *Poblaciones vulnerables ante la salud y el trabajo*. México: El Colegio Mexiquense.
- Navarrete, E. L., Padrón, M. & Silva, A. C. (2014). La inserción laboral de los jóvenes y las políticas de empleo en Colombia, México y Uruguay (2012). En L. Gandini & M. Padrón, *Población y trabajo en América Latina: Abordajes teórico-metodológicos y tendencias empíricas recientes*. Recuperado de [https://www.academia.edu/6049272/Luciana\\_Gandini\\_y\\_Mauricio\\_Padrón\\_Innamorato\\_Coordinadores\\_2014\\_Población\\_y\\_trabajo\\_](https://www.academia.edu/6049272/Luciana_Gandini_y_Mauricio_Padrón_Innamorato_Coordinadores_2014_Población_y_trabajo_)

[en\\_América\\_Latina\\_abordajes\\_teórico-metodológicos\\_y\\_tendencias\\_emp%C3%ADricas\\_recientes\\_](#)

- Negrete, R. (2011). El concepto estadístico de informalidad y su integración bajo el esquema del Grupo de Dehli. *Revista Internacional de Estadística y Geografía*, 2(3). Recuperado de <https://rde.inegi.org.mx/index.php/2011/09/04/el-concepto-estadistico-de-informalidad-y-su-integracion-bajo-el-esquema-del-grupo-de-delhi/>
- Negrete, R. (2012). Sector informal en México visto desde el esquema conceptual OIT-Grupo de Delhi. En E. de la Garza (Coord.) *La situación del trabajo en México, 2012, el trabajo en la crisis* (pp. 119-154). México: Plaza y Valdés. Recuperado de <http://www2.itzt.uam.mx/sotraem/NovedadesEditoriales/Situacion-deltrabajo.pdf>
- Ochoa, S. M. (2016). Trayectorias laborales durante la crisis económica 2008-2009 en México. *Economía Informa*, 399(12), 34-58. <https://doi.org/10.1016/j.ecin.2016.08.004>
- Orozco, K. (2014). *El papel de las cargas domésticas y los arreglos familiares en el trabajo asalariado urbano de México* (Tesis de Doctorado). El Colegio de México. Recuperado de [https://colmex.userservices.exlibrisgroup.com/view/delivery/52COL-MEX\\_INST/1264973180002716](https://colmex.userservices.exlibrisgroup.com/view/delivery/52COL-MEX_INST/1264973180002716)
- Orozco, K. (2015). Participación femenina en trabajos asalariados: ¿una doble selectividad? *Carta Económica Regional*, 17(116). Recuperado de <http://www.cartaeconomicaregional.cucea.udg.mx/index.php/CER/article/view/6142>
- Ortega, A. (2013). Definiendo un Índice Multidimensional de Trabajo Decente para México. *Revista Mexicana de Economía y Finanzas*, 8(1), 75-99. <https://doi.org/10.21919/remef.v8i1.42>
- Ovando, W. & Rodríguez, O. (2013). Flexibilidad laboral y desigualdad salarial. La industria manufacturera mexicana como evidencia, 2005-2010. *Análisis Económico*, 28(67), 59-76.
- Ovando, W., Román, Y. G. & Salgado, M. del C. (2018). Trabajo a tiempo parcial y desigualdad salarial en la industria manufacturera en México (2005-2015). *Ciencia ergo-sum*, 25(3).
- Padrón, M. & Navarrete, E. L. (2012). Una mirada sobre el trabajo infantil en México. El módulo del trabajo infantil de la ENOE. *Coyuntura Demográfica*, 2, 75-80.

- Partida, V. (2011). Estimación indirecta de tasas de ingreso y de retiro de la actividad económica para México. *Estudios Demográficos y Urbanos*, 26(1), 33-73.
- Partida, V. (2014). Cambios en los mercados laborales de México de 2000 a 2010 mediante esperanza de vida. *Papeles de población*, 20(18), 121-164.
- Pedrero, M. (2018). Condiciones precarias de trabajo, una forma de violencia institucional. El caso del Estado de México 2005-2011. *El trabajo y su medición. Mis tiempos. Antología de estudio sobre trabajo y género*. México: Centro Regional de Investigaciones Multidisciplinarias, UNAM; Miguel Ángel Porrúa.
- Pérez, J. A. & Ceballos, G. I. (2019). Dimensionando la precariedad laboral en México de 2005 a 2015, a través del Modelo Logístico Ordinal Generalizado. *Nóesis. Revista de ciencias sociales y humanidades*, 28(55), 109-135. <https://doi.org/10.20983/noesis.2019.1.6>
- Rendall, M. S., Brownell, P. & Kups, S. (2011). Declining return migration from the United States to Mexico in the late-2000s recession: A research note. *Demography*, 48(3), 1049-1058.
- Rodríguez, E. & Velázquez, D. (2014). Empleos, salarios y precarización del salario en México: Una crítica a la propuesta de la reforma laboral. En D. Castro & R. E. Rodríguez (Coords.), *El mercado laboral frente a las transformaciones económicas en México*. (pp. 235-256). Recuperado de <https://www.cise.uaec.mx/downloads/LibrosElectonicos/LibroMercadoLaboral2014.pdf>
- Rodríguez, D., López, B. & Prudencio, D. (2013). *Labor vulnerability and the evolution of the working poor in Mexico*. Recuperado de <http://old.iariw.org/papers/2013/OreggiaPaper.pdf>
- Rodríguez, R. E. & Castro, D. (2014). Análisis de la discriminación salarial por género en Saltillo y Hermosillo: Un estudio comparativo en la industria manufacturera. *Nóesis: Revista de Ciencias Sociales y Humanidades*, 23(46), 80-113.
- Rodríguez, R. E. & Limas, M. (2017). El análisis de las diferencias salariales y discriminación por género por áreas profesionales en México, abordado desde un enfoque regional, 2015. *Estudios Sociales: Revista de Investigación Científica*, 27(49), 121-150.
- Rodríguez, R. E., Ramos, R. & Castro, D. (2018). Brecha salarial por género en los mercados de trabajo público y privado en México

- (2005-2014). *Panorama Económico*, 25(3), 149-172. Recuperado de [https://www.researchgate.net/publication/324991348\\_Brecha\\_salarial\\_por\\_genero\\_en\\_los\\_mercados\\_de\\_trabajo\\_publico\\_y\\_privado\\_en\\_Mexico](https://www.researchgate.net/publication/324991348_Brecha_salarial_por_genero_en_los_mercados_de_trabajo_publico_y_privado_en_Mexico)
- Rodríguez, E. & López, B. (2015). Imputación de ingresos laborales. Una aplicación con encuestas de empleo en México. *El trimestre económico*, 82(325), 117-146.
- Román, Y. G., Montoya, B. J. & Gaxiola , S. C. (2019). Los adultos mayores y su retiro del mercado laboral en México. *Sociedad y Economía*, (37), 87-113. Recuperado de <https://www.redalyc.org/jatsRepo/996/99660265005/99660265005.pdf>
- Román, Y. G. (2017). Jóvenes y sector informal en el Estado de México. Un grupo en desventaja. *Revista Perspectivas Sociales*, 19(2), 41-60. Recuperado de <http://perspectivassociales.uanl.mx/index.php/pers/article/view/14/10>
- Román, Y. G. & Ovando, W. (2016). Flexibilidad laboral de la población ocupada: Un análisis espacial en México, 2005 y 2014. *Sociedad y Economía*, (31), 193-013.
- Román, Y. G. & Sollova, V. (2015). Precariedad laboral de jóvenes asalariados en la ciudad de Toluca, 2005-2010. *Convergencia*, (67), 129-152. Recuperado de <http://www.scielo.org.mx.pbsdi.unam.mx:8080/pdf/conver/v22n67/v22n67a6.pdf>
- Rubio, J. (2017). Sindicalización y precariedad laboral en México. *Región y sociedad*, 29(68), 37-75. Recuperado de <http://www.scielo.org.mx/pdf/regsoc/v29n68/1870-3925-regsoc-29-68-00037.pdf>
- Ruiz, P. & Ordaz, J. L. (2011). Evolución reciente del empleo y el desempleo en México. *Economía UNAM*, 8(23), 91-105.
- Salas, I. A. (2018). Análisis de las trayectorias laborales en México desde la perspectiva de la calidad del empleo. *Nova Scientia*, 10(21), 576-604.
- Salas, I. A. & Murillo, F. (2013). Los profesionistas universitarios y el mercado laboral mexicano: Convergencias y asimetrías. *Revista de la educación superior*, 42(165), 63-81.
- Samaniego, B. (2017). *Formal firms, informal workers, and household labor supply in Mexico*. EUA: University of Chicago. Recuperado de <https://knowledge.uchicago.edu/record/832>
- Sánchez, A., Villarespe, V., Román, D. A. & Herrera, A. L. (2016).

- Determinantes de las horas de trabajo de las mujeres en México: Un enfoque de pseudopanel (2005-2010). *Revista de la CEPAL*, 120, 127-139.
- Torres, A. J. & Félix, G. (2017). El uso de internet y su relación con los salarios en México: Un análisis no paramétrico. En G. L. Ochoa & A. J. Torres, *Los retos del cambio económico actual: Revisión y aplicaciones para el caso mexicano* (pp. 31-50). Recuperado de [https://www.researchgate.net/publication/322644687\\_El\\_uso\\_de\\_Internet\\_y\\_su\\_relacion\\_con\\_los\\_salarios\\_en\\_Mexico\\_un\\_analisis\\_no\\_parametrico?enrichId=rgreq-59bfa228cb8cf6b3cc652fbe4985766a-XXX&enrichSource=Y292ZXJQYWdlOzMyMjY-0NDY4NztBUzo1ODU3MTAzNjA4ODcyOTZAMTUxNjY1NTY0MDAxMg%3D%3D&el=1\\_x\\_3&\\_esc=publicationCoverPdf](https://www.researchgate.net/publication/322644687_El_uso_de_Internet_y_su_relacion_con_los_salarios_en_Mexico_un_analisis_no_parametrico?enrichId=rgreq-59bfa228cb8cf6b3cc652fbe4985766a-XXX&enrichSource=Y292ZXJQYWdlOzMyMjY-0NDY4NztBUzo1ODU3MTAzNjA4ODcyOTZAMTUxNjY1NTY0MDAxMg%3D%3D&el=1_x_3&_esc=publicationCoverPdf)
- Torres, J. (2013). *Size-dependent firm regulations and the return to skill: Evidence from the Mexican labor market* (Doctor of Philosophy, University of Chicago). Recuperado de <https://search-proquest-com.pbidi.unam.mx:2443/pqdtglobal/docview/1446690100/fulltextPDF/AAE2A2D4EC944E67P-Q/1?accountid=14598>
- Trejo, J. C., Rivera, E. C. & Ríos, H. (2017). Analysis of the hysteresis of unemployment in Mexico in the face of macroeconomic shocks. *Contaduría y Administración*, 62(4), 1249-1269. <https://doi.org/10.1016/j.cya.2017.06.013>
- Understanding Children's Work (UCW) Programme (2012). *The Mexican experience in reducing child labour: Empirical evidence and policy lessons*. Recuperado de World Bank website: <http://documentos.bancomundial.org/curated/es/456481541739883791/pdf/131904-WP-PUBLIC-ADD-SERIES-See-73756-Report-Child-Labour-trends20130308-111116.pdf>
- Varela, R. (2015). Income Differences on Mexico's Northern Border: A Perspective on Formal and Informal Employment. *Frontera Norte*, 27(53), 177-203.
- Varela, R. & Retamoza, A. (2011). Los salarios en México: Un análisis con datos de panel. *Investigación y Ciencia*, 19(53), 29-38.
- Varela, R. & Retamoza, A. (2012). Capital humano y diferencias salariales en México, 2000-2009. *Estudios Fronterizos*, 13(26), 175-200.

Villarreal, A. (2014). Explaining the Decline in Mexico-U.S. Migration: The Effect of the Great Recession. *Demography*, 51(6), 2203-2228.

Villarreal, A. & Blanchard, S. (2013). How Job Characteristics Affect International Migration: The Role of Informality in Mexico. *Demography*, 50(2), 751-775.

## Código de análisis

La base de datos se puede descargar del enlace <<https://tinyurl.com/Cap1-Literatura>>.

```
#### Capítulo I - Análisis de palabras clave ####
### Settings

rm(list=ls()) ### borra objetos en el ambiente
setwd("~/") ### establece el directorio donde descargaste la base de datos

if (!require("pacman")) install.packages("pacman") # instala pacman si se requiere

## Loading required package: pacman

#cargan Los paquetes necesarios. No se utilizan todos
pacman::p_load(tidyverse, tm, NLP, SnowballC, wordcloud2, ggwordcloud,
                 quanteda, e1071, plyr, stringr, MASS, sentimentr,
                 sjlabelled, knitr, jsonlite, rjson, RColorBrewer, XML,
                 reshape2, ggraph, qgraph, topicmodels, readxl, janitor,
                 broom, ggthemes, wesanderson, ggwordcloud)

##### Importando la base #####
gral <- read_excel("Revisión LiteraturaV1.xlsx", sheet = "gral")
gral<-clean_names(gral)

gral_noautor<-gral %>%
  group_by(correlativo) %>%
  add_tally() %% #agrega el número de autores
  ungroup() %%
  mutate(no_autor=n)

gral_noautor$palabras_clave<-tolower(gral_noautor$palabras_clave)
gral_noautor$titulo<-tolower(gral_noautor$titulo)
gral_noautor$resumen<-tolower(gral_noautor$resumen)
gral_noautor$perspectiva_2<-tolower(gral_noautor$perspectiva_2)
gral_noautor$metodo<-tolower(gral_noautor$metodo)

#Nos vamos a quedar con la info del autor principal
gral_noautor<-gral_noautor %>%
  filter(lugar_autor==1 & year>2009 & year<2020)

gral_noautor %>%
  tabyl(idioma)

##   idioma   n percent
##   Español 102  0.6375
```

```

##   Inglés 58 0.3625

gral_noautor %>%
  tabyl(sex, idioma) %>%
  adorn_totals(c("col", "row")) %>%
  adorn_percentages("col")

##     sex Español Inglés Total
##     F 0.4411765 0.29310345 0.38750
##     M 0.5588235 0.68965517 0.66625
## <NA> 0.0000000 0.01724138 0.00625
## Total 1.0000000 1.0000000 1.00000

gral_noautor %>%
  summarise(promedio=mean(no_autor))

##   promedio
## 1 1.94375

summary(gral_noautor$no_autor)

##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
## 1.000 1.000 2.000 1.944 3.000 5.000

gral %>%
  tabyl(sex)

##     sex n percent valid_percent
##     F 130 0.40000000 0.4012346
##     M 194 0.596923077 0.5987654
## <NA> 1 0.003076923 NA

gral_noautor %>%
  tabyl(perspectiva_2)

##           perspectiva_2 n percent
##           ccss 14 0.08750
##           desarrollo 9 0.05625
##           economía 72 0.45000
##           estadística 4 0.02500
##           estudios laborales 6 0.03750
##   estudios latinoamericanos 2 0.01250
##           otra 11 0.06875
##           políticas públicas 3 0.01875
##           regional 8 0.05000
##           sociodemográfica 25 0.15625
##           sociología 6 0.03750

gral_noautor %>%
  tabyl(perspectiva_2, idioma) %>%
  adorn_totals(c("col", "row")) %>%
  adorn_percentages("col")

##           perspectiva_2 Español Inglés Total
##           ccss 0.107843137 0.05172414 0.08750
##           desarrollo 0.019607843 0.12068966 0.05625
##           economía 0.362745098 0.60344828 0.45000
##           estadística 0.029411765 0.01724138 0.02500
##           estudios laborales 0.049019608 0.01724138 0.03750
##   estudios latinoamericanos 0.009803922 0.01724138 0.01250
##           otra 0.088235294 0.03448276 0.06875
##           políticas públicas 0.019607843 0.01724138 0.01875
##           regional 0.058823529 0.03448276 0.05000
##           sociodemográfica 0.205882353 0.06896552 0.15625

```

```
#### Numeralia ####

gral_noautor %>%
  ggplot(aes(x=as.factor(year))) + geom_bar()

pal2 <- wes_palette("Darjeeling1", 4)

gral_noautor %>%
  ggplot(aes(x=as.factor(year), fill=sex)) + geom_bar() + facet_wrap(~idioma) +
  scale_fill_manual(labels = c("Mujeres", "Hombres", "NA"),
                     values=pal2)+ labs(x="Año", y="# Textos") + theme_minimal()

ggsave(plot=last_plot(), filename="C1_g1.jpeg", width = 8, height =5)

pal3 <- wes_palette("Darjeeling1", n=11, type="continuous")

gral_noautor %>%
  ggplot(aes(x=perspectiva_2, fill=perspectiva_2)) + geom_bar() + facet_wrap(~idioma) +
  scale_fill_manual(values=pal3)+ labs(x="Disciplina", y="# Textos") +
  theme_minimal() + theme(legend.position = "none") + coord_flip()

ggsave(plot=last_plot(), filename="C1_g3.jpeg", width = 8, height =5)

gral_noautor %>%
  filter(!is.na(sex)) %>%
  ggplot(aes(x=perspectiva_2, fill=perspectiva_2)) + geom_bar() + facet_wrap(~sex) +
  scale_fill_manual(values=pal3)+ labs(x="Disciplina", y="# Textos") +
  theme_minimal() + theme(legend.position = "none") + coord_flip()

ggsave(plot=last_plot(), filename="C1_g3b.jpeg", width = 8, height =5)

#### tokenización ####

library(udpipe)
library(lattice)

### Bajando los modelos
udmodel <- udpipe_download_model(language = "spanish") # esto trabaja con la estructura del ESPAÑOL

udmodel_eng <- udpipe_download_model(language = "english") # esto trabaja con la estructura del inglés

## Downloading udpipe model from https://raw.githubusercontent.com/jwijffels/udpipe.models.ud.2.4/master/inst/udpipe-ud-2.4-190531/english-ewt-ud-2.4-190531.udpipe to /Users/Libro/Cap1/english-ewt-ud-2.4-190531.udpipe
## Visit https://github.com/jwijffels/udpipe.models.ud.2.4 for model license details

#### Títulos ####
x_title_spa <- udpipe(x = paste(unique(gral_noautor[gral_noautor$idioma=="Español",]$titulo)), object=udmodel)
x_title_eng <- udpipe(x = paste(unique(gral_noautor[gral_noautor$idioma=="Inglés",]$titulo)), object=udmodel_eng)

#### Palabras clave ####

x_key_spa<- udpipe(x = paste(unique(gral_noautor[gral_noautor$idioma=="Español" & !is.na(gral_noautor$palabras_clave),]$palabras_clave)), object=udmodel)
x_key_eng<- udpipe(x = paste(unique(gral_noautor[gral_noautor$idioma=="Inglés" & !is.na(gral_noautor$palabras_clave),]$palabras_clave)), object=udmodel_eng)
```

```

##### métodos #####
x_met_spa<- udpipe(x = paste(unique(gral_noautor$metodo)), object=udmodel)

##### Resúmenes #####
xx_spa<-paste(unique(gral_noautor[gral_noautor$idioma=="Español",]$resumen))
xx_eng<-paste(unique(gral_noautor[gral_noautor$idioma=="Inglés",]$resumen))

x_spa <- udpipe(x = xx_spa, object=udmodel)
x_eng <- udpipe(x = xx_eng, object=udmodel_eng)

##### sustantivos de palabras claves #####
### Español
statsk_spa<- subset(x_key_spa, upos %in% "NOUN")
statsk_spa <- txt_freq(statsk_spa$token)
statsk_spa$key <- factor(statsk_spa$key, levels = rev(statsk_spa$key))
barchart(key ~ freq, data = head(statsk_spa, 20), col = "cadetblue",
         main = "Los sustantivos más usados [Español]", xlab = "Frecuencia")

statsk_spa$idioma<-"Español"

### Inglés
statsk_eng<- subset(x_key_eng, upos %in% "NOUN")
statsk_eng <- txt_freq(statsk_eng$token)
statsk_eng$key <- factor(statsk_eng$key, levels = rev(statsk_eng$key))
barchart(key ~ freq, data = head(statsk_eng, 20), col = "cadetblue",
         main = "Los sustantivos más usados [Inglés]", xlab = "Frecuencia")

statsk_eng$idioma<-"Inglés"

## Gráfico 2
rbind(statsk_spa[1:20,], statsk_eng[1:20,]) %>%
  ggplot(aes(x=key, y=freq)) + geom_bar(stat="identity") +
  geom_text(aes(label=freq), vjust=-0.3, size=3.5) +
  coord_flip() + facet_wrap(~idioma, scales="free") +
  scale_fill_manual(values=pal2)+ labs(y="Frecuencia", x="Sustantivo") + theme_minimal()

ggsave(plot=last_plot(), filename="C1_g2.jpeg", width = 8, height = 5)

##### Resúmenes - sustantivo #####
xx_eng<-as.data.frame(xx_eng)
xx_eng$row<-rownames(xx_eng)
xx_eng$doc_id<-paste("doc",xx_eng$row, sep="")
names(xx_eng)<-c("resumen", "row", "doc_id")

x_eng<-merge(x_eng, xx_eng, by="doc_id", all=T)
x_eng<-merge(x_eng, gral_noautor, by="resumen", all.x = T)

xx_spa<-as.data.frame(xx_spa)
xx_spa$row<-rownames(xx_spa)
xx_spa$doc_id<-paste("doc",xx_spa$row, sep="")
names(xx_spa)<-c("resumen", "row", "doc_id")

x_spa<-merge(x_spa, xx_spa, by="doc_id", all=T)
x_spa<-merge(x_spa, gral_noautor, by="resumen", all.x = T)

```

```
#palabras muy comunes
comunes<-c("trabajo", "empleo", "objetivo", "documento", "méxico", "mexico", "resultados",
"trabajadores", "ocupación",
"presente", "artículo", "estudio", "parte", "partir", "datos", "encuesta", "enoé",
"mercado")
stop<-c(comunes,stopwords("spanish"))

## NOUNS
res_spa<-x_spa %>%
  filter(!token %in% stop) %>%
  filter(upos %in% "NOUN") %>%
  with(txt_freq(token))

res_spa$key <- factor(res_spa$key, levels = rev(res_spa$key))

res_spa$idíoma<-"Español"

commons<-c("employment", "workers", "mexico", "survey", "data", "paper", "document",
"results", "labor", "sector")
stop2<-c(stopwords("english"), commons)

## NOUNS
res_eng<-x_eng %>%
  filter(!token %in% stop2) %>%
  filter(upos %in% "NOUN") %>%
  with(txt_freq(token))

res_eng$idíoma<-"Inglés"
res_eng$key <- factor(res_eng$key, levels = rev(res_eng$key))

rbind(res_spa[1:20,], res_eng[1:20,]) %>%
  ggplot(aes(x=key, y=freq)) + geom_bar(stat="identity") +
  geom_text(aes(label=freq), vjust=-0.3, size=3.5) +
  coord_flip() + facet_wrap(~idioma, scales="free") +
  scale_fill_manual(values=pal2)+ labs(y="Frecuencia", x="Sustantivo") + theme_minimal()

ggsave(plot=last_plot(), filename="C1_gres.jpeg", width = 8, height =5)

#### Resúmenes - rake #####
rake_spa <- keywords_rake(x = x_spa,
                           term = "token", group = c("doc_id", "paragraph_id", "sentence_id"),
                           relevant = x_spa$upos %in% c("NOUN", "ADJ"),
                           ngram_max = 5)
head(subset(rake_spa), 10)

##                 keyword ngram freq      rake
## 1  conocimientos administrativos generales   3    2 5.500000
## 2          mínimos cuadrados ordinarios   3    2 5.333333
## 3            capital humano informal   3    2 3.422436
## 4             mercado laboral mexicano   3    4 3.405459
## 5        mayor participación femenina   3    2 2.947884
## 6           política económica   2    2 2.439560
## 7            américa latina   2    2 2.333333
## 8            distrito federal   2    3 2.333333
## 9        industria manufacturera   2    5 2.309524
## 10       sector manufacturero   2    2 2.298851

rake_eng <- keywords_rake(x = x_eng,
                           term = "token", group = c("doc_id", "paragraph_id", "sentence_id"),
                           relevant = x_eng$upos %in% c("NOUN", "ADJ"),
                           ngram_max = 5)
head(subset(rake_eng),10)
```

```

##                  keyword ngram freq      rake
## 1      universal health care    3   3 5.910714
## 2      mexican labor market    3   6 4.625916
## 3      labor market duality   3   2 4.444963
## 4      social security       2   2 4.431818
## 5      urban wage disparities  3   2 4.155914
## 6      global financial crisis 3   2 3.934615
## 7      national unemployment rates 3   2 3.578283
## 8      family income inequality 3   2 3.534188
## 9      formal sector job      3   3 3.532857
## 10     human capital          2   2 3.500000

openxlsx::write.xlsx(rake_spa, "rake_spa.xlsx")
openxlsx::write.xlsx(rake_eng, "rake_eng.xlsx")

##### Métodos - rake #####
rake_met <- keywords_rake(x = x_met_spa,
                           term = "token", group = c("doc_id", "paragraph_id", "sentence_id"),
                           relevant = x_met_spa$upos %in% c("NOUN", "ADJ"),
                           ngram_max = 5)

head(subset(rake_met),10)

##                  keyword ngram freq      rake
## 1 regresión logística multinomial 3   2 4.404762
## 2             regresión lineal   2   2 2.821429
## 3      regresión logística   2   4 2.404762
## 4      análisis descriptivo   2   5 2.083333
## 5      efectos aleatorios   2   2 2.000000
## 6      efectos fijos        2   2 2.000000
## 7      variables instrumentales 2   2 2.000000
## 8      análisis comparativo   2   2 1.916667
## 9      modelo económetrico   2   2 1.838235
## 10     equilibrio general    2   2 1.666667

openxlsx::write.xlsx(rake_met, "rake_met.xlsx")

```

**¿Cómo empezar a estudiar el mercado de trabajo en México?  
Una introducción al análisis estadístico con R aplicado a la  
Encuesta Nacional de Ocupación y Empleo**

La primera edición digital realizada por  
la Facultad de Ciencias Políticas y Sociales  
de la Universidad Nacional Autónoma de México,  
se finalizó el 27 de agosto de 2021.

Diseño y formación de portada e interiores:  
iGIRA, en calle Esfuerzo s/n, Plan de Ayala,  
C. P. 14470, Alcaldía Tlalpan, Ciudad de México.

En la composición de la obra se utilizaron las familias tipográficas  
Times New Roman y Hevetica Neue en 11/13 pts.

Cuidado editorial:  
Departamento de Publicaciones, FCPyS, UNAM