Redacción de artículos académicos con R

Episodio 8: Gráficos estadísticos con tidyverse

Bajaña Alex

Chanatasig Evelyn

Heredia Aracely

2022-07-16


1
2
3
4
5
6

Librería y datos

Librería y datos

```
library(tidyverse)

source("../redaccion_documento/scripts/mod_base_tid.R", encoding = "UTF-8")

# La base final se llama "enemdu", la renombramos
tabla <- enemdu
```

Inferencia estadística para ciencia de datos

Variables aleatorias

Variable aleatoria

> 🚨En el episodio 2 hablamos por primera vez de las variables aleatorias y su relación con los vectores en R.

Deciamos que una variable aleatoria es una función que asigna un valor, usualmente numérico, al resultado de un experimento aleatorio.

Una variable aleatoria es la colección de los posibles resultados de un experimento en una población dada. En el caso de lanzar una moneda, el experimento es lanzar una moneda, la variable aleatoria es cada resultado: cara y sello.

La inferencia estadística se trata de describir poblaciones utilizando datos. Es importante recordar que de todo lo hablaremos es una cantidad de población, no una declaración sobre lo que ocurre en nuestros datos. Por ejemplo, al lanzar una moneda, el hecho de que el 50% de probabilidad de que salga cara es una afirmación sobre la moneda y cómo la lanzamos, no una afirmación sobre el porcentaje de caras que obtuvimos en una serie particular de lanzamientos.

Tipos y ejemplos

Las variables aleatorias que estudiamos vendrán en dos variedades, discretas o continuas:

- Las variables aleatorias discretas son variables aleatorias que solo aceptan un número contable de posibilidades. Las funciones de probabilidad asignarán probabilidades de que tomen valores específicos.
- La variable aleatoria continua puede tomar conceptualmente cualquier valor en la línea real o algún subconjunto de la línea real y hablamos de la probabilidad de que se encuentren dentro de algún rango. Las densidades caracterizarán estas probabilidades.

Ejemplo de variables aleatorias discretas. Los experimentos de juego familiares, como el lanzamiento de una moneda y el lanzamiento de un dado, producen variables aleatorias. Para la moneda, normalmente codificamos una cola como un 0 y una cara como un 1. (Para el dado, el número hacia arriba sería la variable aleatoria).

Ejemplo de variables aleatorias continuas. En cosas como longitudes o pesos. Es matemáticamente conveniente modelarlos como si fueran continuos (incluso si las mediciones se truncaron generosamente).

Ejemplo de experimento

Lanzar una moneda: En este ejemplo se lanzarán 100 veces una moneda.

En la siguiente tabla se muestran los resultados del experimento. Por ejemplo, el resultado del primer lanzamiento fue cara, mientras que el resultado del centésimo lanzamiento fue sello.

lanzamiento	resultado
x(1)	cara
x(2)	sello
x(3)	cara
x(4)	cara
x(5)	cara
...	...
x(100)	sello

Ejemplo de experimento

En este ejemplo, la variable `lanzamiento` es una **variable aleatoria**. Para hacer este ejercicio en R se usa la función `sample()`,

```
lanzamiento <- sample(c("cara", "sello"), 100, replace=T)
```

lanzamiento
sello
cara
sello
cara
cara
cara
sello

Cada vez que se ejecuta el comando, aparecen valores distindos, esto es porque no se ha declarado una semilla, si en R no se detecta una semilla, la variable aleatoria resultante va a tener distindos valores. Si no se declara una semilla es como si se realizara este experimento con 100 personas diferentes cada vez.

Más ejemplos de variables aleatorias

- El número de carros que pasan por el kilómetro 10 de la Avenida Panamericana norte en una hora, de 17:00 a 18:00. La hora es importante para eliminar los casos externos como las horas pico. Esta variable aleatoria sigue una **distribución de Poisson**.
- El tiempo de espera en un peaje para que el automovil pase cuando se está en vehículo particular. Esta variable sigue una **distribución exponencial**.
- Al jugar "pichirilo qué color" en un viaje. Dependiendo de la ruta que se tome, la probabilidad de los valores cercanos a cero serán más probables en viajes largos. Esta variable tiene **Distribución Binomial**.

> Las distribuciones mencionadas se explicarán más adelante.

Ejemplo de experimento

Caso ENEMDU: En este caso, el experimento es el levantamiento de la información y la variable aleatoria son cada una de las variables como: los rubros de ingreso, entre otros. La información fue recolectada de acuerdo a las condiciones del INEC. Enlace a ANDA del INEC

Las condiciones del diseño tienen gran importancia puesto que al realizarlo mal, la información recolectada puede estar sesgada.

Bajo el supuesto de que podamos ver al mismo individuo en 2 períodos de la encuesta, se podría medir la probabilidad de que una persona pierda su empleo. En este caso la función de probabilidad podría tener dos opciones: logística y de bernoulli.

Función de probabilidad:

1. Densidad de Probabilidad
2. Distribución Acumulada

Función de probabilidad

Una función de probabilidad evaluada en un valor corresponde a la probabilidad de que una variable aleatoria tome ese valor. Para ser una función válida, debe satisfacer:

1. Siempre debe ser mayor o igual a 0.
2. La suma de las probabilidades asociadas a los posibles valores que puede tomar la variable aleatoria tiene que sumar uno.

Funciones de densidad de probabilidad (fdp)

Una función de densidad de probabilidad (fdp), es una función asociada con una variable aleatoria continua. Las áreas bajo las funciones de densidad de probabilidad corresponden a probabilidades para esa variable aleatoria.

Por ejemplo, cuando dicen que los cocientes de inteligencia (CI) en la población siguen una curva de campana, están diciendo que la probabilidad de que una persona seleccionada al azar de esta población tenga un CI entre dos valores viene dada por el área bajo la curva de campana.

Funciones de densidad de probabilidad (fdp)

No todas las funciones pueden ser una función de densidad de probabilidad válida. Por ejemplo, si la función cae por debajo de cero, entonces podríamos tener probabilidades negativas. Si la función contiene demasiada área debajo de ella, podríamos tener probabilidades mayores que uno. Las siguientes dos reglas nos dicen cuándo una función es una función de densidad de probabilidad válida. Específicamente, para ser un fdp válido, una función debe satisfacer:

- Debe ser mayor o igual a cero en todas partes.
- El área total debajo de él debe ser uno.

Función de Distribución Acumulada (FDA)

La función de distribución acumulativa (FDA) de una variable aleatoria, X, devuelve la probabilidad de que la variable aleatoria sea menor o igual que el valor x.

Esta definición se aplica independientemente de si la variable aleatoria es discreta o continua.

Tipos de Funciones de Distribución

Lo que vamos a ver en esta sección se cubrió de forma breve en la clase 3, en las diapositivas de la 6 a la 8.

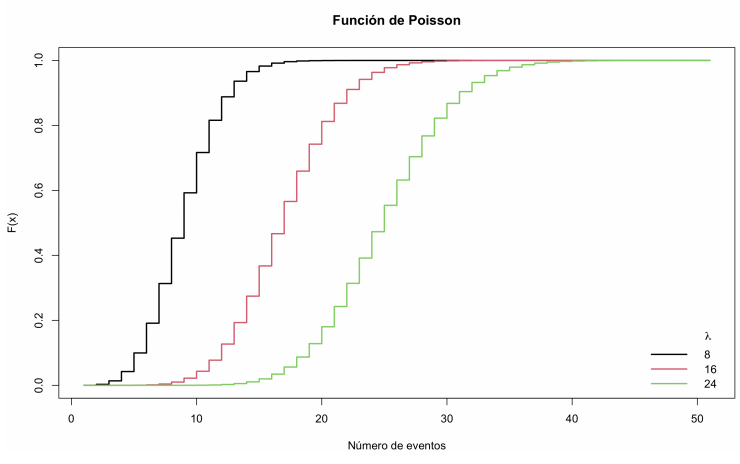
Función de Poisson

Función de Poisson

La distribución de **Poisson** es discreta y modela el número de veces que ocurre un evento en un intervalo de tiempo. El parámetro λ representa el número de veces que se espera que ocurra el fenómeno durante un intervalo dado.

Se aplica a varios fenómenos discretos de la naturaleza que ocurren `0, 1, 2, 3, ...`, veces **durante un periodo definido de tiempo** o en un área determinada. Ejemplos de eventos que pueden ser modelados con Poisson incluyen:

- Número de errores ortográficos que comete al escribir una única página
- Número de llamadas telefónicas en una central telefónica por minuto



- Número de estrellas en un determinado volumen de espacio

En el gráfico se observa que en el eje x la función solamente está definida en valores enteros . Las líneas que conectan los vértices son guías visuales y no indican continuidad.

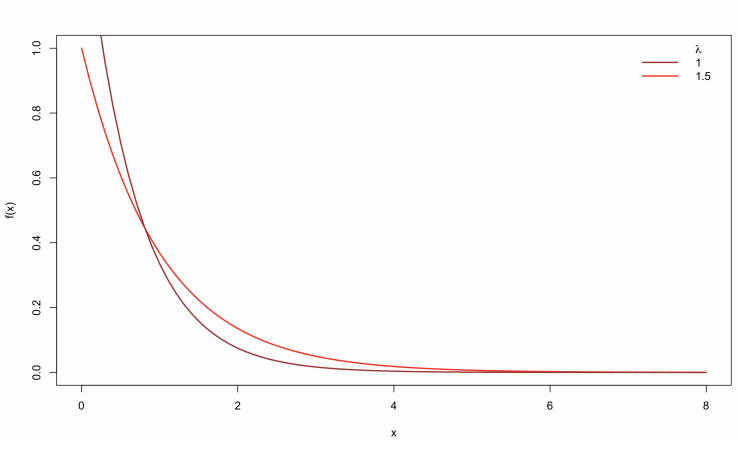
Distribución Exponencial

Distribución Exponencial

La distribución de **Exponencial** es una distribución continua que se utiliza para modelar tiempos de espera para la ocurrencia de un cierto evento.

Describe procesos en los que interesa saber el tiempo hasta que ocurre determinado evento. Un ejemplo es el tiempo que tarda una partícula radiactiva en desintegrarse.

La **distribución exponencial** se puede caracterizar como la distribución del tiempo entre sucesos consecutivos generados por un proceso de Poisson; por ejemplo, el tiempo que transcurre entre dos heridas graves sufridas por una persona.



La media de la distribución de Poisson,λ, que representa la **tasa de ocurrencia del evento por unidad de tiempo, es el parámetro de la distribución exponencial**, y su inversa es el valor medio de la distribución.

Distribución Exponencial

Ejemplos para la distribución exponencial es la distribución de la longitud de los intervalos de una variable continua que transcurren entre dos sucesos, que se distribuyen según la distribución de Poisson:

- El tiempo transcurrido en un centro de llamadas hasta recibir la primera llamada del día se podría modelar como una exponencial.
- El intervalo de tiempo entre terremotos (de una determinada magnitud) sigue una distribución exponencial.
- Supongamos una máquina que produce hilo de alambre, la cantidad de metros de alambre hasta encontrar una falla en el alambre se podría modelar como una exponencial.

Distribución Binomial

Distribución Binomial

Una distribución binomial es una distribución de probabilidad discreta que describe el número de éxitos al realizar n experimentos independientes entre sí, acerca de una variable aleatoria.

Cuenta el número de éxitos en una secuencia de n ensayos de Bernoulli independientes entre sí con una probabilidad fija p de ocurrencia de éxito entre los ensayos.

Ejemplo: En un viaje a Cayambe, existen cuatro personas jugando "pichirilo qué color", por el total de número de carros que ven, sólo unos pocos serán pichirilos. El número total de carros representa a la función de Bernoulli y el éxito de pichirilos representa a la función binomial.

Distribución Binomial

Experimento

El número total de carros representa a la función de Bernoulli y el éxito de pichirilos representa a la función binomial.

Persona A

Al desglosar el número de éxitos de la persona A, los resultados se verían más o menos como en la tabla anterior.

> n éxitos en N intentos.

La probabilidad condicional

Motivación

El condicionamiento es un tema central en estadística. Si se nos da información sobre una variable aleatoria, cambia las probabilidades asociadas a ella.

Por ejemplo, la probabilidad de obtener uno al lanzar un dado (estándar) generalmente se asume que es un sexto, en el caso en el que **el dado no esté alterado**. Ahora imaginemos que Pedro y Juan están jugando con los dados y, Pedro recibirá $1 cada vez que el dado dé un número impar (por lo tanto, 1, 3 o 5), para Pedro, la probabilidad de un uno es ahora un tercio.

Considere otro ejemplo. Con el resultado de una prueba de diagnóstico por imágenes para el cáncer de pulmón. ¿Cuál es la probabilidad de que una persona tenga cáncer si la prueba es positiva? ¿Cómo cambia esa probabilidad sabiendo que un paciente ha sido un fumador empedernido de por vida y ambos padres tenían cáncer de pulmón? Dependiendo de esta nueva información, la probabilidad ha aumentado drásticamente.

Definición

Probabilidad condicional es la probabilidad de que ocurra un evento `A`, sabiendo que también sucede otro evento `B`. Matemáticamente:

Sea `B` un evento de modo que `P(B) > 0`. Entonces, la probabilidad condicional de un evento `A` dado que `B` ha ocurrido es (y se lee «la probabilidad de `A` dado `B`».):

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

Si `A` y `B` no están relacionados de ninguna manera, es decir, son independientes, entonces:

$$P(A \mid B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

Si la ocurrencia de `B` no ofrece información sobre la ocurrencia de `A`, la probabilidad condicionada a la información es la misma que la probabilidad sin la información, en este caso, decimos que los dos eventos son independientes.

Graficar funciones

Función de densidad

Gráficar una función de densidad

Cuando se va a graficar una función de densidad se usa la función `geom_density()`, en esta se pueden especificar ciertas características como:

- `alpha = ...:` se especifica la opacidad que se desea tener en el gráfico para que la superposición de las densidades no resulte un problema en la visualización del gráfico.
- `position = 'stack':` separa las densidades, mmuestra la distribución apilada por categorías

Fig 1

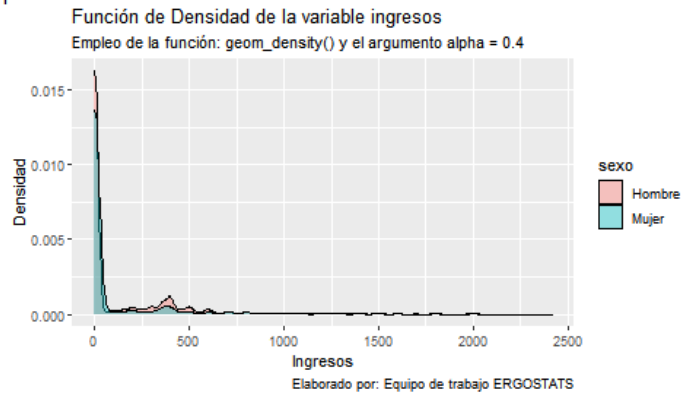
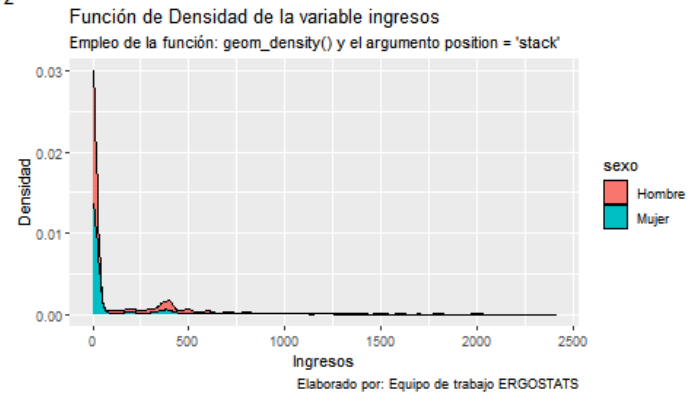


Fig 2



Distribución de densidad por grupos

Fig 3

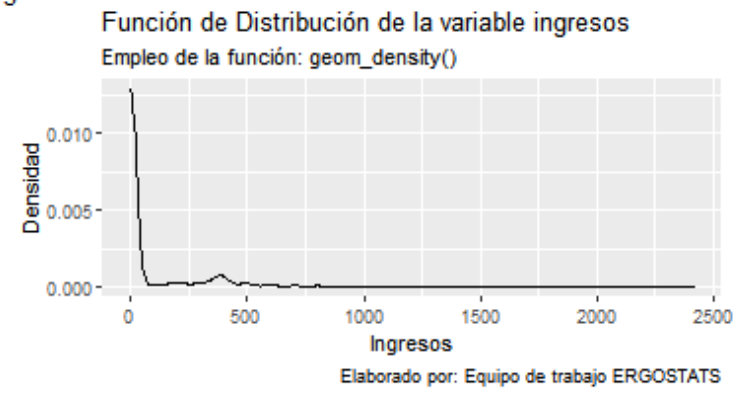


Fig 4

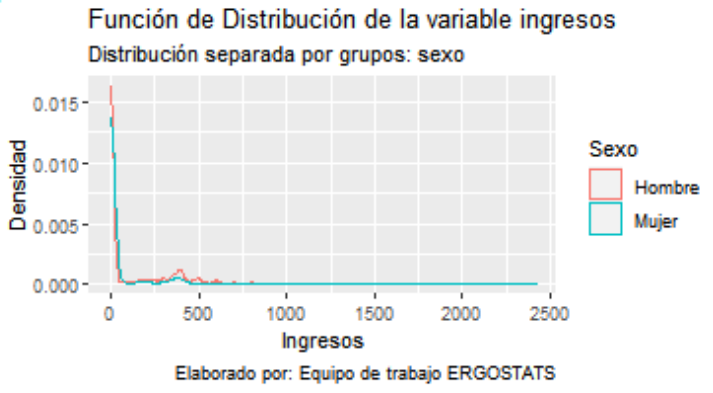


Fig 5

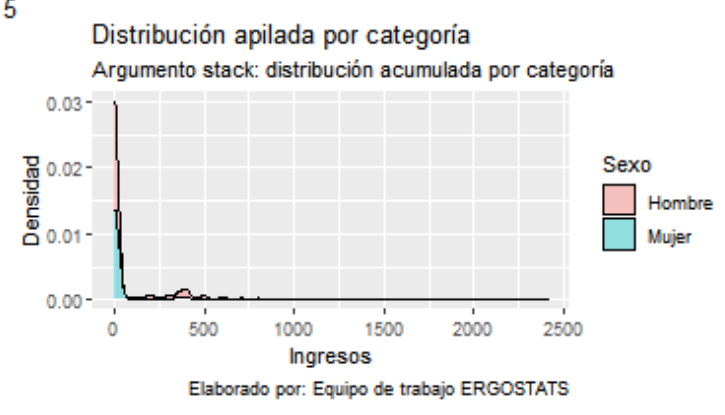
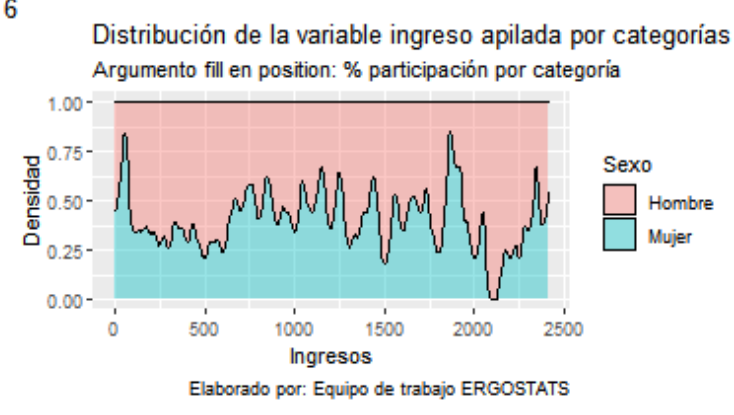


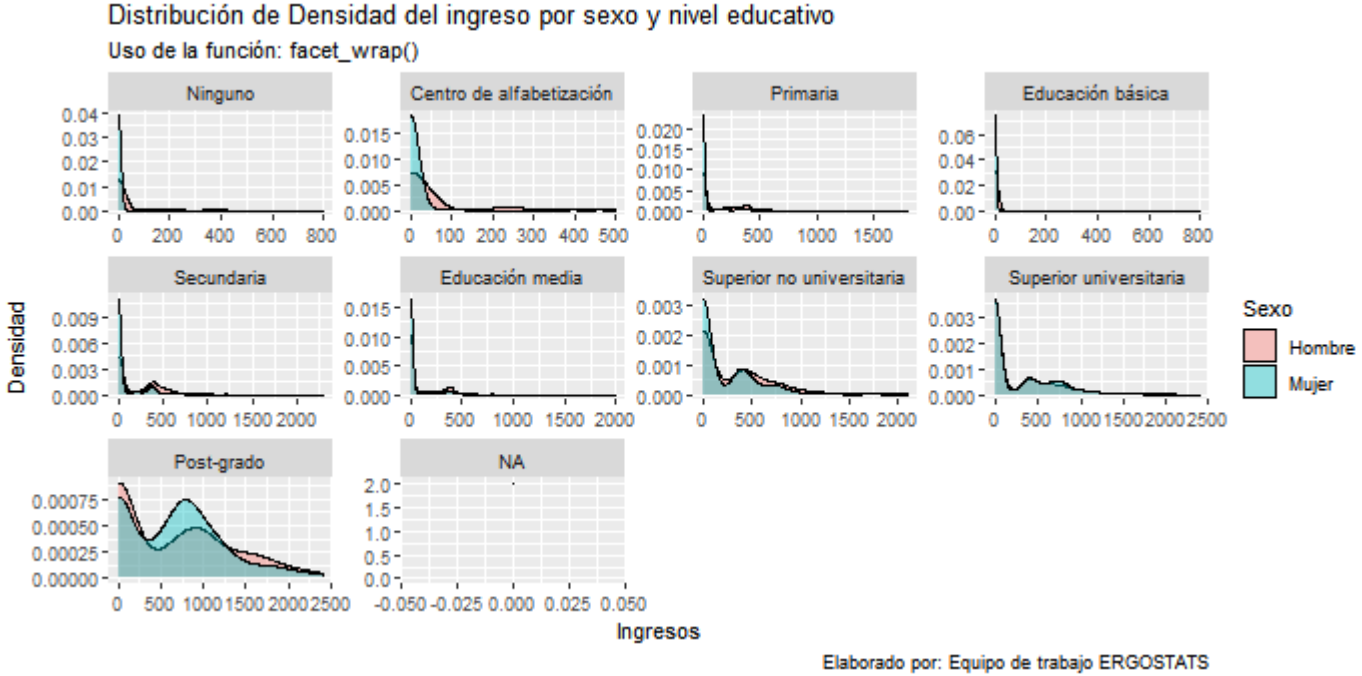
Fig 6



Con `facet_wrap()`

En este ejemplo se hace el mismo ejemplo anterior, con la diferencia de que se usa la función `facet_wrap()`. Como es de conocimiento, esta función se emplea para visualizar una variable categórica de forma independiente por cada grupo que existe.

Fig 7



Elaborado por: Equipo de trabajo ERGOSTATS

Gráfico de líneas

Gráfico de líneas: código

```
tabla %>%
  group_by(h_job,sexo) %>%
  summarise(ing_promedio = mean(ing,na.rm = T)) %>%
  ggplot(mapping = aes(x = h_job,
                       y = ing_promedio,
                       color = sexo)) +
  geom_line() +
  labs(title = "Gráfico de líneas entre ingresos promedio y horas de trabajo a la semana",
       subtitle = "Uso del ggplot para ver la relación entre variables",
       caption = "Elaborado por: Equipo de trabajo ERGOSTATS",
       tag = "Fig. 8",
       x = "Horas de Trabajo a la semana", y = "Ingresos promedio", color="Sexo")+
  scale_x_continuous(n.breaks = 30, limits = c(0,87))  +
  scale_y_continuous(n.breaks = 10, limits = c(0, 1090)) +
  geom_text(x = 4, y = 1070,
            label = "Nos llamó la \n atención este pico",
            show.legend = F)+
  geom_text(x = 43, y = 1050,
            label = "40 horas de trabajo a la semana",
            show.legend = F)
```

Gráfico de líneas: funciones extras

Como los gráficos de líneas muestran una serie como un conjunto de puntos conectados mediante una sola línea, entonces, se deben tener valores únicos por cada grupo en la variable `x`.

Como es un gráfico de líneas y, en este caso existen picos. Para etiquetar o poner un comentario en estos puntos, se usa la función `geom_text()`, aquí se especifica las coordenadas y el texto que se quiere incluir. `show_leyend=F` se coloca para que, el texto escrito no se muestre en la leyenda.

En la función `scale_x_continuous()`y `scale_y_continuous()` se puede especificar argumentos adicionales a los ya vistos, como son:

- `n.break:` añade puntos al eje para que tenga una mejor visualización
- `limits:` especifica el espacio visual de los ejes

Gráfico de líneas: gráfico

Fig. 8

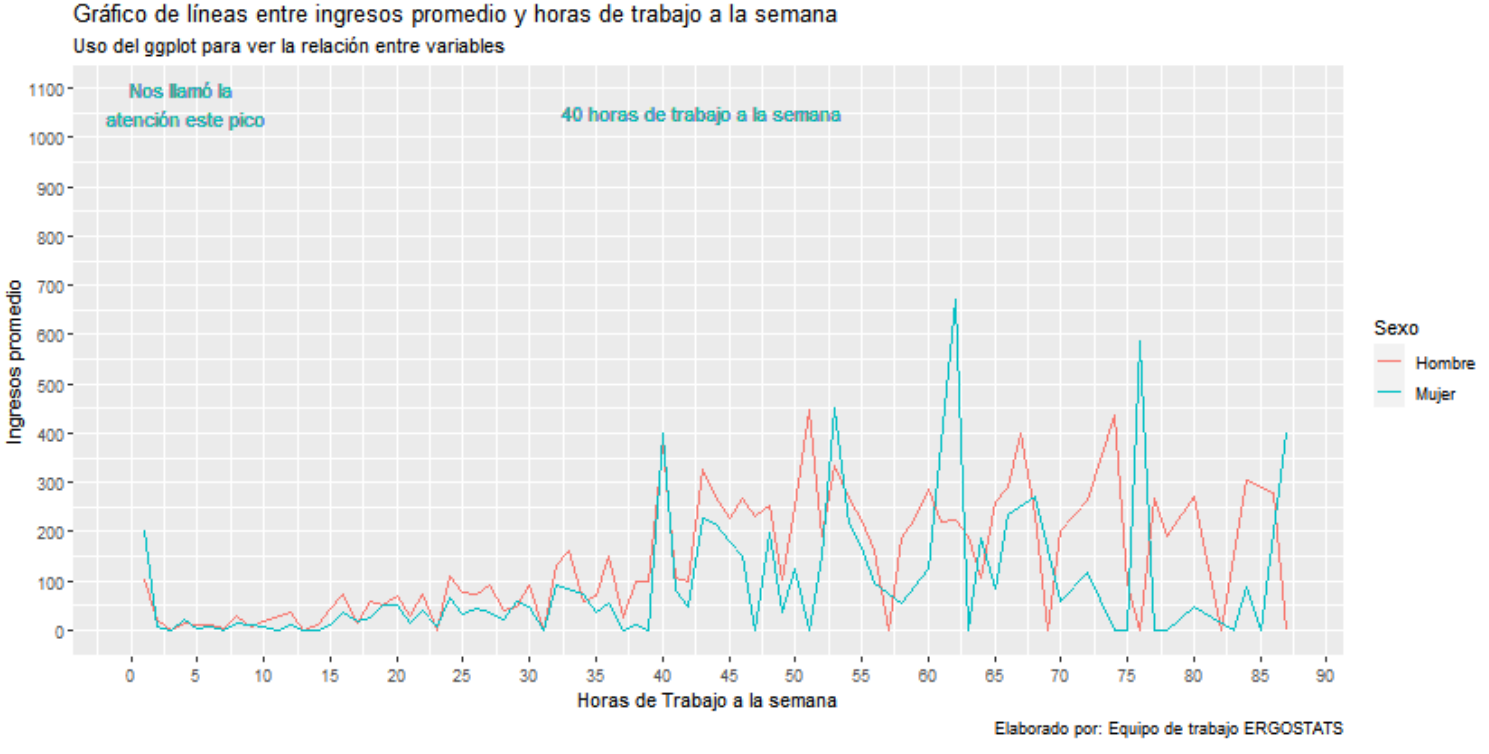


Gráfico de Cajas:
`geom_boxplot()`

Gráfico de Cajas: explicación

Para crear un gráfico de cajas se usa la función `geom_boxplot`, dentro de ente en el argumento `aes()` se especifica las variables que van en el eje `x`, eje `y`, y, la variable categórica por la cual se definirá el `color`.

Gráfico de Cajas: gráfico

En el siguiente gráfico de cajas se usó la variable `ingresos`. La explicación de este gráfico es la siguiente:

La caja de un `boxplot` comienza en el primer cuartil (25%) y termina en el tercero (75%). Por lo tanto, la caja representa el 50% de los datos centrales, con una línea dentro que representa la mediana. A cada lado de la caja se dibuja un segmento con los datos más lejanos sin contar los valores atípicos (outliers) del box plot, que en caso de existir, se representarán con círculos.

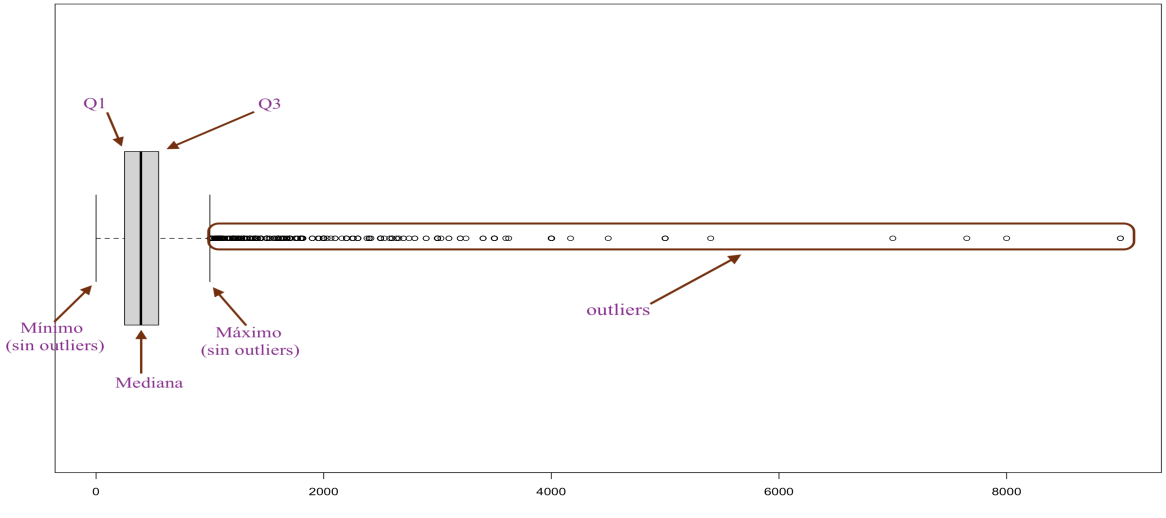
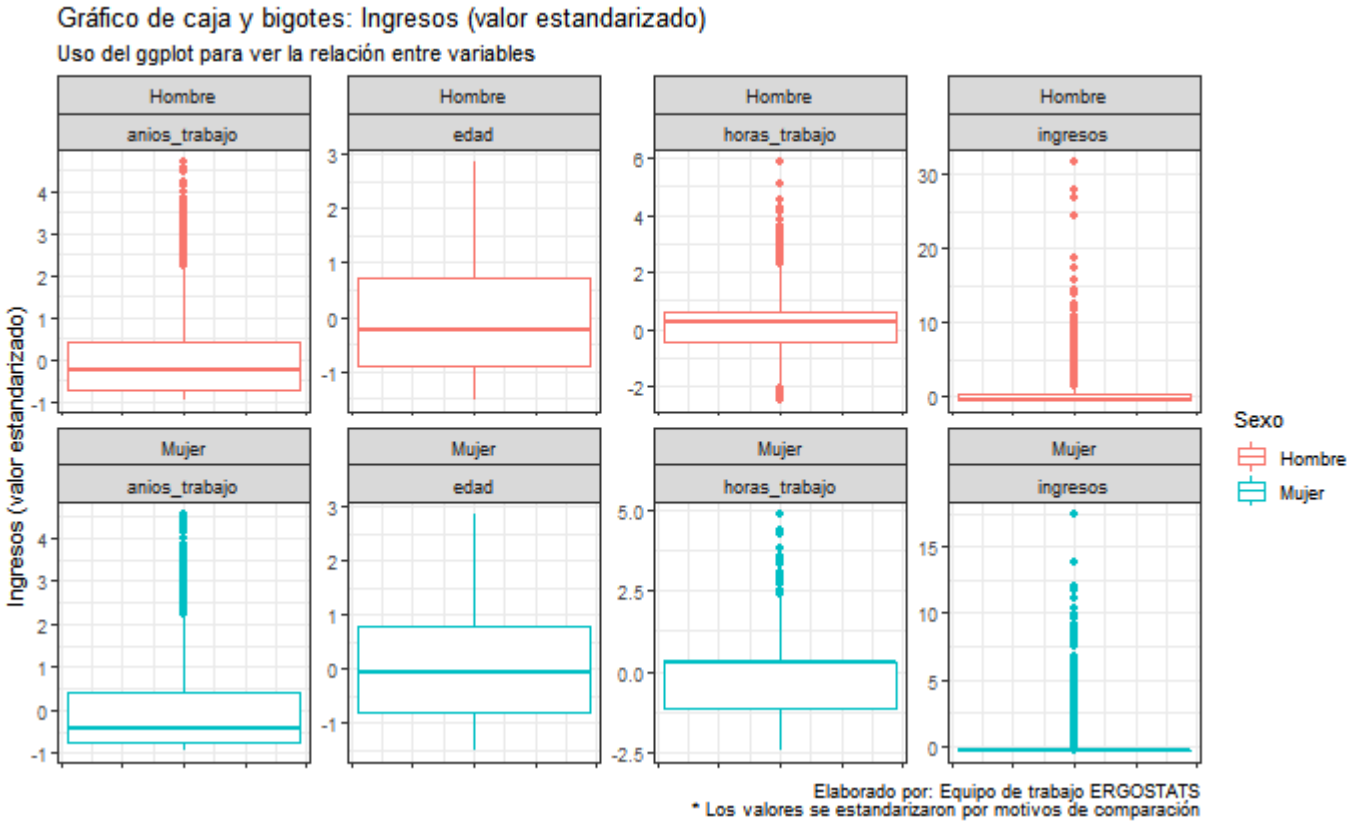


Gráfico de Cajas: código

```
tabla %>%
  select(sexo,edad,a_job,h_job,ing) %>%
  rename(anios_trabajo= a_job,
         horas_trabajo=h_job,
         ingresos=ing)%>%
  mutate_if(is.numeric,scale) %>%
  as_tibble %>%
  gather(variable,valor,-sexo) %>%
  ggplot() +
  geom_boxplot(aes(x = 1,
                   y = valor,
                   group = variable,
                   color = sexo))+
  facet_wrap(sexo ~ variable,scales = "free",ncol = 4,nrow = 2) +
  labs(title = "Gráfico de caja y bigotes: Ingresos (valor estandarizado)",
       subtitle = "Uso del ggplot para ver la relación entre variables",
       caption = "Elaborado por: Equipo de trabajo ERGOSTATS\n* Los valores se estandariza
       tag = "Fig. 9",
       x = "Variables", y = "Ingresos (valor estandarizado)",
       color = "Sexo")+
  theme_bw()+
  theme(axis.text.x = element_blank(), axis.title.x = element_blank())
```

Gráfico de Cajas: gráfico

Fig. 9



Histograma

Gráficar usando histogramas

Cuando se va a graficar un histograma se usa la función `geom_histogram()`, en esta se pueden especificar ciertas características como:

- `position = "dodge"` separa las barras por categoría
- `bins` aumenta el número de barras
- `binwidth` se especifica el ancho de la barra en términos de la variable

Ejemplos:

```
tabla %>% filter(ing<2500) %>%
  ggplot(mapping = aes(x = ing, fill = sex
  geom_histogram()
```

```
tabla %>% filter(ing<2500) %>%
  ggplot(mapping = aes(x = ing, fill = sex
  geom_histogram(position = "dodge")
```

```
tabla %>% filter(ing<2500) %>%
  ggplot(mapping = aes(x = ing, fill = sex
  geom_histogram(bins = 50, position = "do
```

```
tabla %>% filter(ing<2500) %>%
  ggplot(mapping = aes(x = ing, fill = sex
  geom_histogram(binwidth = 50, position =
```

Gráfico usando histogramas

Fig. 10

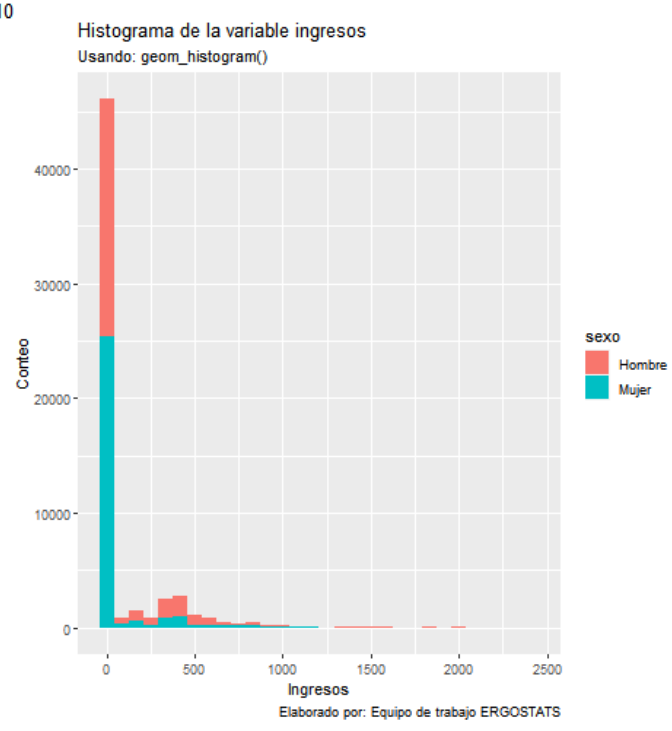


Fig. 11

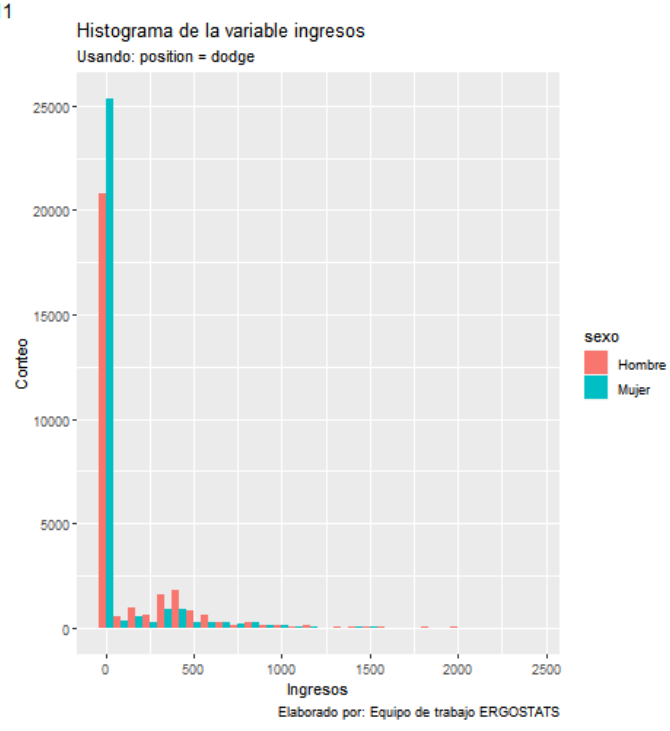


Gráfico usando histogramas

Fig. 12

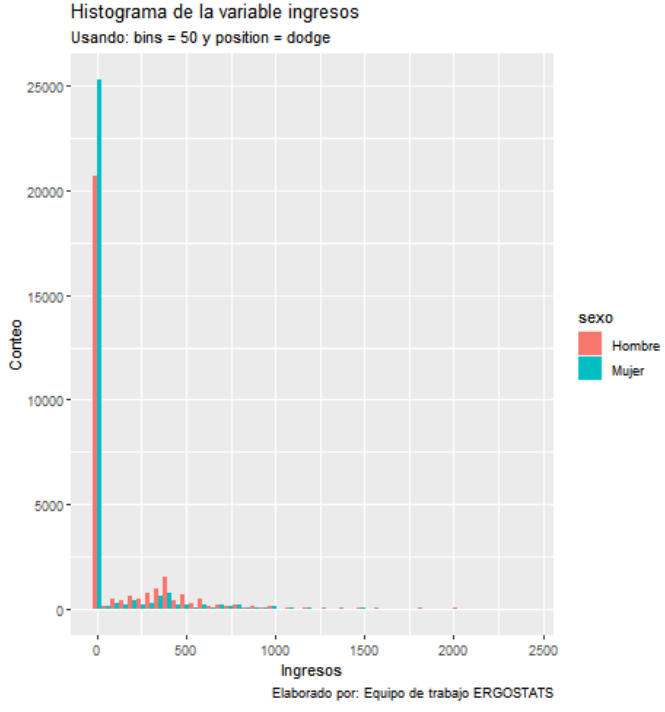


Fig. 13

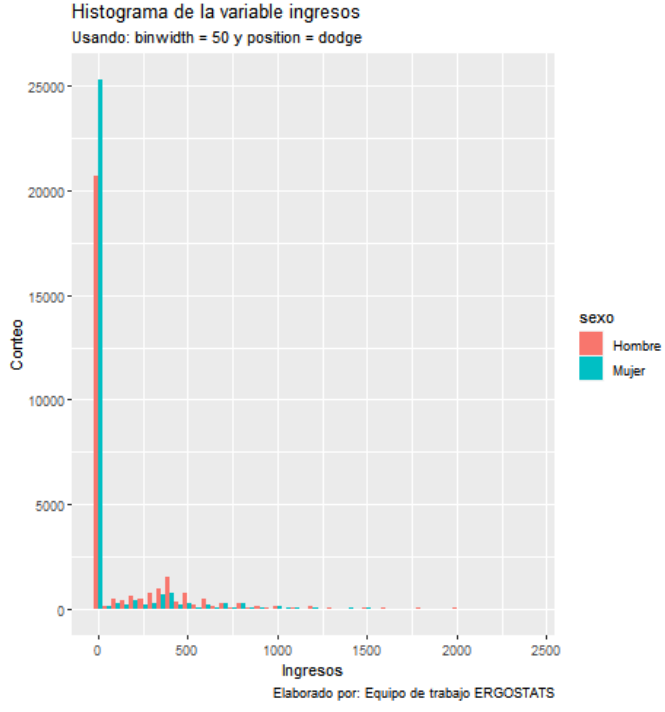
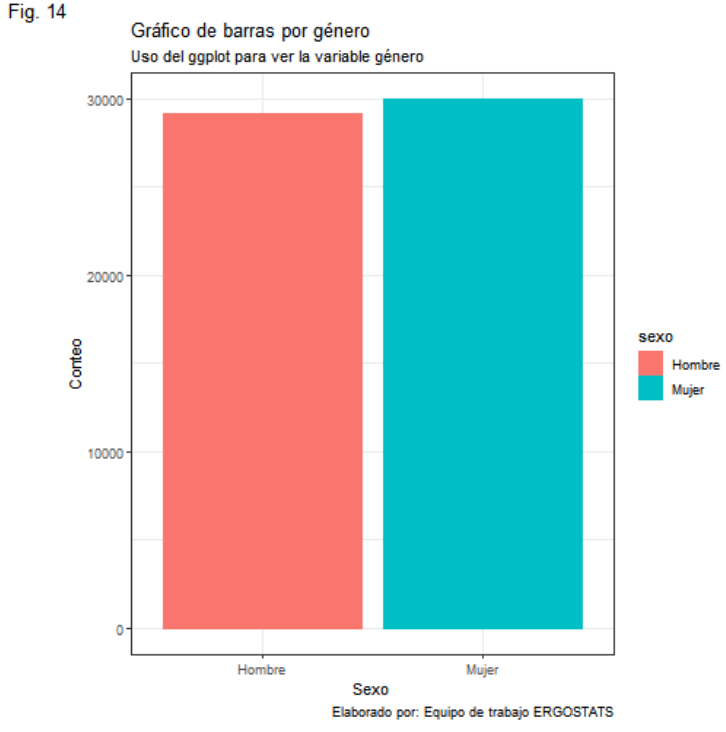


Gráfico de barras

Con una variable

`geom_bar()` está diseñado para facilitar la creación de gráficos de barras que muestren recuentos o sumas

```
tabla %>% ggplot(mapping = aes(x=sexo, col
  labs(title = "Gráfico de barras por géne
       subtitle = "Uso del ggplot para ver
       caption = "Elaborado por: Equipo de
       tag = "Fig. 14",
       x = "Sexo", y = "Conteo")+
  theme_bw()
```

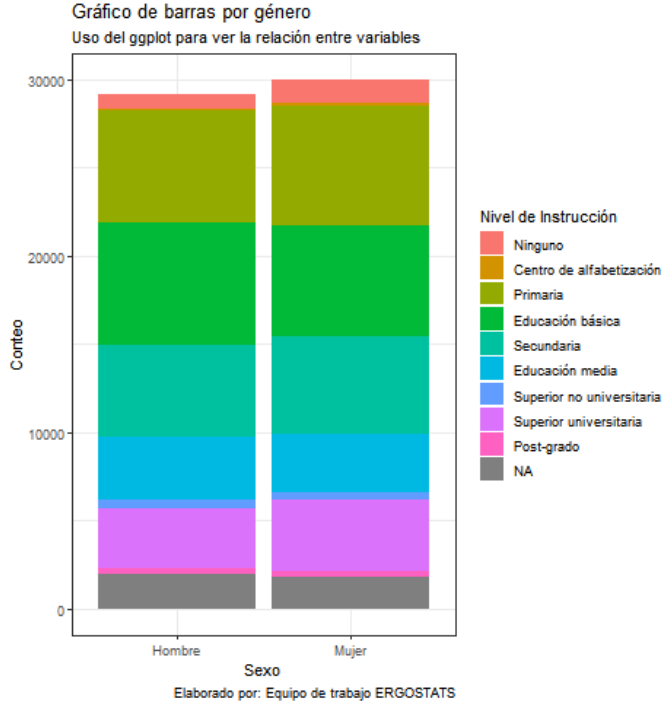


Con dos variables

En este gráfico se muestra el conteo de la variable `sexo` mientras que el relleno del color de las barras está clasificado por `nivel de instrucción`.

```
tabla %>% ggplot(mapping = aes(x=sexo, fil
  geom_bar()+
  labs(title = "Gráfico de barras por géne
       subtitle = "Uso del ggplot para ver
       caption = "Elaborado por: Equipo de
       tag = "Fig. 15",
       x="Sexo", y = "Conteo", fill= "Nive
  theme_bw()
```

Fig. 15

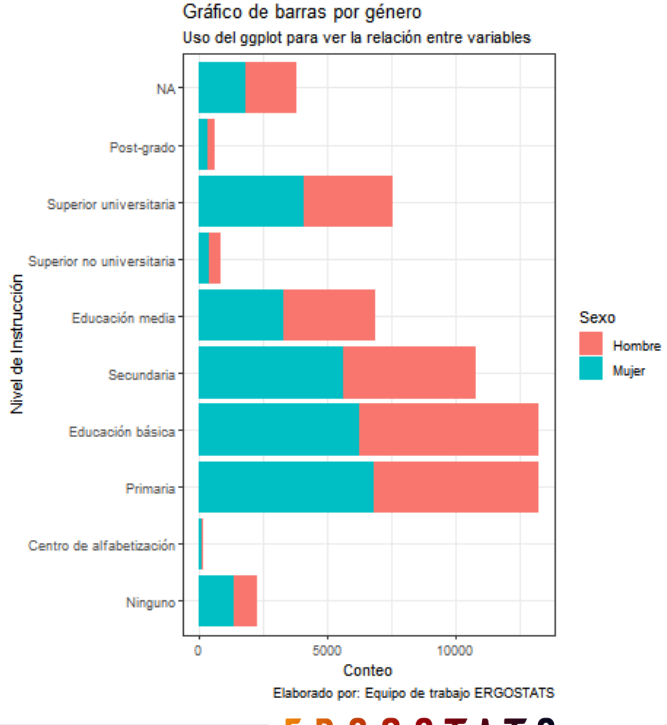


Con dos variables

Este gráfico muestra la misma información que el anterior, la diferencia es que en el eje `y` se ubicó a la variable `nivel de instrucción` mientras que se rellenaron las barras con la variable `sexo`.

```
tabla %>% ggplot(mapping = aes(y=niv_inst,
    geom_bar()+
    labs(title = "Gráfico de barras por gé
         subtitle = "Uso del ggplot para v
         caption = "Elaborado por: Equipo
         tag = "Fig. 16",
         x="Conteo", y = "Nivel de Instruc
    theme_bw()
```

Fig. 16



Características

Para hacer una gráfica de barras apilada porcentual se usa el argumento `position="fill"` dentro de la función `geom_bar()`. En este caso se representa el porcentaje de cada subgrupo, lo que permite estudiar la evolución de su proporción en el conjunto.

Fig. 17

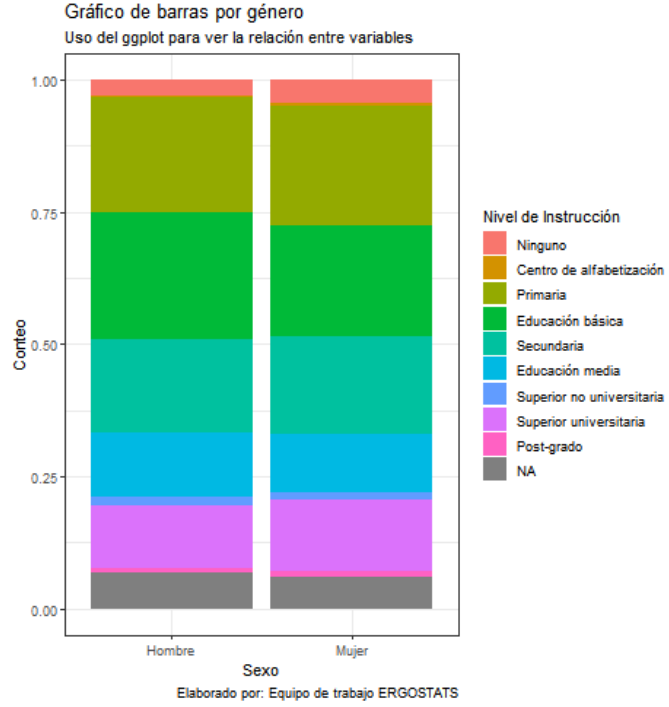
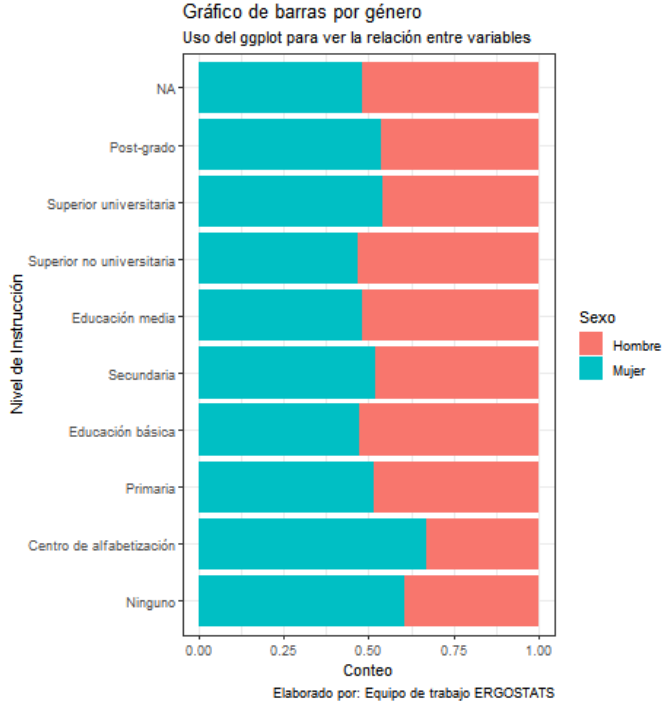


Fig. 18



Gráficos con Normas APA séptima edición

Componentes de una figura

Las figuras de estilo APA tienen los siguientes componentes básicos:

- **número de la figura**: (por ejemplo, Figura 1) es el primer ítem que debemos agregar. Se debe usar negrita. Se enumeran las figuras en el orden en que aparecen en el documento.
- **título**: el título de la figura debe aparece una línea debajo del número de la figura. Es necesario dar a cada figura un título breve pero descriptivo. Utilice cursiva en el título.
- **imagen**: se refiere al gráfico, fotografía, dibujo u otra ilustración.
- **leyenda**: debe ser colocada dentro de los bordes de la figura y puede ser usada para explicar los símbolos utilizados en la imagen de la misma.
- **nota**: aquí se agrega cualquier contenido que se necesite describir que no pueden entenderse solo por el título o por la imagen por si misma (por ejemplo, definiciones de abreviaturas, atribución de derechos de autor). Se incluye notas de figuras solo si es necesario.

Ubicación de las figuras en el texto

Hay dos opciones para la ubicación de figuras (y tablas) en una investigación. Se puede incluir cada figura en el texto después de que la menciones por primera vez, se puede agregar cada figura en una página separada después de la lista de referencias, o después de las tablas (si las hay).

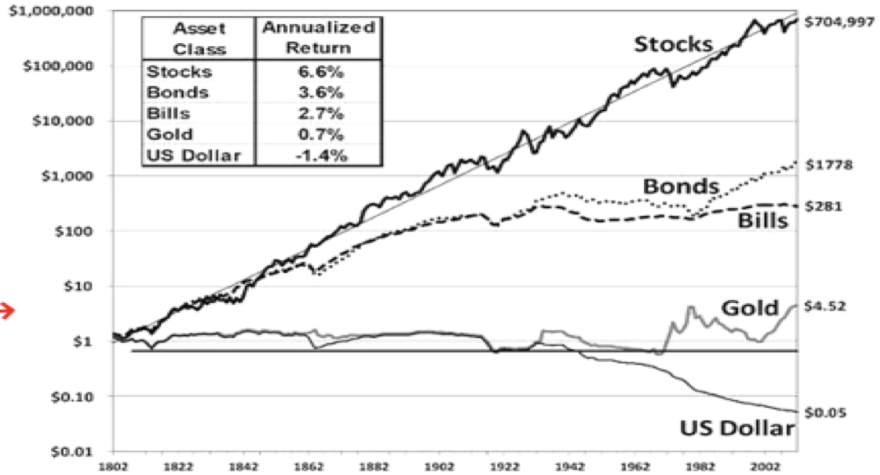
Ejemplo

Figura 1 ← número de la figura

título de la figura ←

Retorno real de acciones americanas, títulos del tesoro americano, oro y dólar de 1802 a 2012

imagen →



Nota. El gráfico representa el retorno descontado de la inflación en el período, por eso, un dólar en 2012 vale menos que un dólar en 1802. Tomado de *Stocks for the Long Run* (p.120), por J. J. Siegel, 2014, McGrawHillEducation.

nota ↖

Enlace para saber más acerca de figuras normas APA

Modificar el tamaño en Rmarkdown

Hay una serie de opciones que afectan la salida de figuras dentro de los documentos:

- **fig.height**: se usa para controlar el alto de la figura, y
- **fig.width**: se usa para controlar el ancho.

Por defecto las medidas son: 6.5 x 4.5

Estas opciones se especifican dentro del chunk, a continuación se muestra un ejemplo:

```
```{r fig.height=3, fig.width=7}

código de la figura

```
```

Modificar el tamaño de todos los gráficos

También se puede modificar el tamaño de todos los gráfico en el `yaml` para que todos tengan las mismas dimensiones, para ello se lo define de la siguiente manera:

```yaml
---
title: "Habits"
output:
  pdf_document:
    fig_width: 7
    fig_height: 6
    fig_caption: true
---
```


THE
END

Recursos

- Statistical inference for data science
- Easy multi-panel plots in R using facet_wrap() and facet_grid() from ggplot2
- Normas APA séptima edición
- Modificar tamaño de gráficos
- Distribuciones de Probabilidad