# Taller T.O UCH. Diciembre 2021

El taller tiene como objetivo introducir un software de análisis estadístico, realizar un análisis de datos e interpretar los resultados de nuestro análisis. La base de datos utilizada corresponde a ENDISC-II. Nuestras variables dependientes serán los determinantes sociales de sexo, edad y educación, mientras que nuestra variable dependiente será discapacidad.

*Web ENDISC-II:*

https://www.senadis.gob.cl/pag/355/1197/ii_estudio_nacional_de_discapacidad

*Web RStudio:*

https://www.rstudio.com/products/rstudio/download/#download

*Instalación y primeros pasos*

https://conceptosclaros.com/instalar-r-primeros-pasos/

*GitHub Taller*

https://github.com/eloluna/to_uch_dic21

*Web con el producto del taller*

https://rpubs.com/eloluna/840182

## Intro R

- Intro RStudio interfaz y script
- Paquetes, data y comandos
- Comentarios
- Markdown, chunks y outputs

### Paquetes

### R Notebook

Dos partes, un procesador de texto y 'chunks' para escribir comandos. Lo escrito aquí es el **procesador**. Los chunks son aquellas secciones de otro color que comienzan con "'{r}

### Comentarios

En los chunks, todo lo que lleve "#" va a ser interpretado por R como un comentario y no lo va a correr como comando.

```
#este es un comentario
#esto me va a generar problemas si no tiene #
```

# Taller

Realizar análisis descriptivos de ENDISC-II. Recordemos que nuestras variables dependientes serán los determinantes sociales de sexo, edad, educación e ingresos, mientras que nuestra variable dependiente será discapacidad.

## Cargar datos y examinar

Una vez que tengo una idea de lo que pretendo realizar, es necesario cargar los datos y examinar su estructura.

## Limpieza

Si reviso la metodología de la encuesta, tendré una idea de por qué existe tan alto porcentaje de NAs (missing data) en algunas variables. Quien encuesta entrevista a una persona y esta da información acerca del grupo familiar, pero ciertas variables, como discapacidad, solo son respondidas por el encuestado.

Ejemplo: El encuestado vive en una casa con su pareja, su suegra y sus dos hijos. Quien realiza la entrevista va a solicitar información acerca de elementos como sexo, edad, ocupación, entre otros para todos los miembros del hogar (5 personas), sin embargo, para variables más complejas, como las relacionadas con discapacidad, solo se le pedirá al encuestado.

Esto se vería así en una base:

```
id_hogar <- c(1,1,1,1,1)

id_persona <- c(1,2,3,4,5)

var_sexo <- c(0,1,1,0,1) #0=Hombre, 1=Mujer

var_edad <- c(55, 50, 72, 20, 17)

var_occ <- c("Empleado", "Empleado", "Jubilado",
             "Estudiante", "Estudiante")

var_disc <- c(25, NA, NA, NA, NA) #Solo un dato

tibble(id_hogar, id_persona, var_sexo,
       var_edad, var_occ, var_disc)
```

```
## # A tibble: 5 x 6
##   id_hogar id_persona var_sexo var_edad var_occ    var_disc
##      <dbl>      <dbl>    <dbl>    <dbl> <chr>         <dbl>
## 1        1          1        0       55 Empleado         25
## 2        1          2        1       50 Empleado         NA
## 3        1          3        1       72 Jubilado         NA
## 4        1          4        0       20 Estudiante       NA
## 5        1          5        1       17 Estudiante       NA
```

Cuál es el riesgo de no tomar en cuenta aquello?

```
#Promedio de edad en toda la muestra
mean(endisc$edad)
```

```
## [1] 36.95154
```

```
#Promedio de edad de personas con mi variable de interés
endisc2 <- subset(endisc, !is.na(endisc$cap_puntaje_adulto))

mean(endisc2$edad)
```

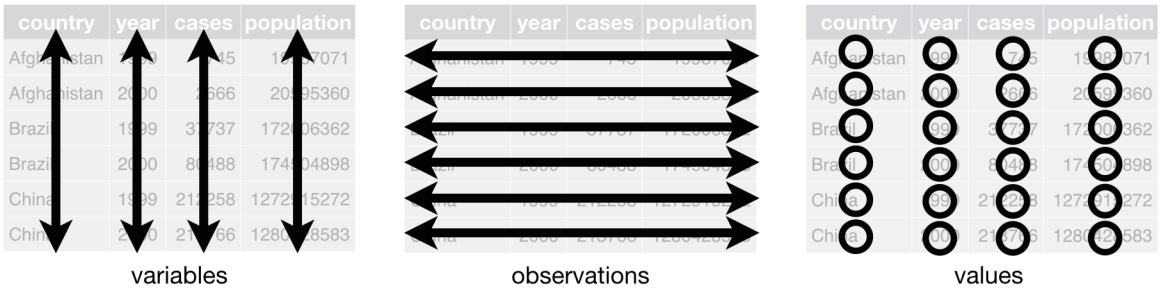
```
## [1] 48.35907
```

```
#Diferencias en nuestra estimación. Necesitamos tener un número determinado que sea constante para todos
#complete.cases(endisc)
summary(complete.cases(endisc))
```

```
##    Mode   FALSE    TRUE
## logical   27637   12249
```

```
endisc_cc <- subset(endisc, complete.cases(endisc))
#endisc_cc
```

*Nota sobre estructura de datos*




## Etiquetado

Ver data antes y después

```
##
##    1    2
## 5303 6946
```

```
## Hombre  Mujer
##      1      2
```

```
##
## Hombre  Mujer    Sum
##   5303   6946  12249
```

```
##
##    0    1    2    3    4    5    6   99
## 4010 8655 3266 6409 8584 3802 5115   45
```

```
## Sin educación formal    Básica incompleta     Básica completa.
##                    0                    1                    2
##     Media incompleto       Media completo  Superior incompleta
##                    3                    4                    5
##    Superior completa             Sin dato
##                    6                   99
```

```
##
## Sin educacion formal    Basica incompleta      Basica completa
##                  337                 1906                 1414
##     Media incompleta       Media completa  Superior incompleta
##                 1760                 3410                 1290
##    Superior completa                  Sum
##                 2132                12249
```

```
##
##    0    1    2
## 9633 1527 1089
```

```
##             Sin Discapacidad Discapacidad Leve a Moderada
##                            0                           1
##          Discapacidad Severa
##                            2
```

```
##
## Sin Discapacidad  Leve a Moderada           Severa              Sum
##             9633             1527             1089            12249
```

## Descripción

Vamos a realizar descripciones de nuestras variables de interés: discapacidad, sexo, edad y educación. Qué método utilizo para describir estas variables depende de su naturaleza.

### Sexo

La variable sexo es una variable binaria, por lo tanto, podemos describirla a través de frecuencias.

**Tabla 1. Número y porcentaje de personas en la muestra por sexo**

| Sexo | N | % |
|---|---|---|
| Hombre | 5,303 | 43.29 |
| Mujer | 6,946 | 56.71 |

### Edad

La variable edad es continua. Podemos examinar sus medidas centrales y de dispersión, y generar un histograma para examinar su distribución.

```
#summary(endisc_cc$edad)
#Medidas de tendencia central

#Promedio
round(mean(endisc_cc$edad, na.rm = T),2)
```

```
## [1] 48.36
```

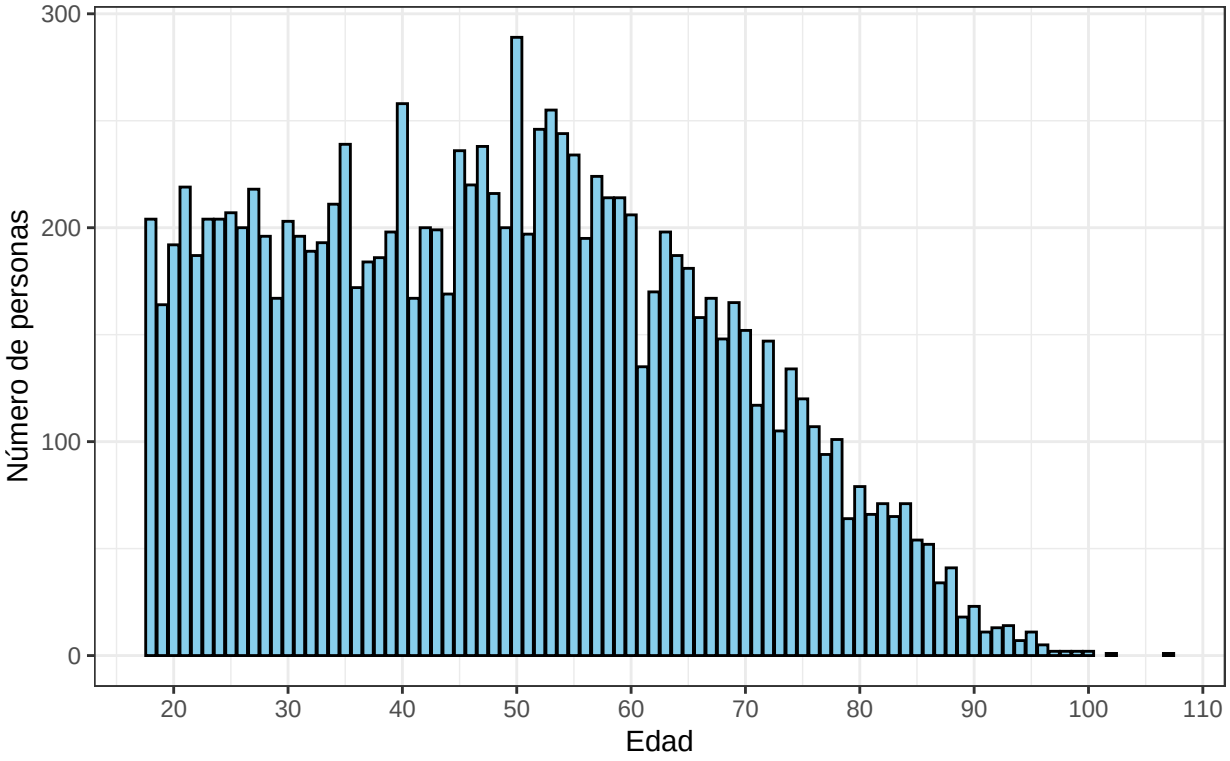
```
#Mediana
median(endisc_cc$edad, na.rm = T)
```

```
## [1] 48
```

```
#Desviación stándard
sd(endisc_cc$edad, na.rm = T)
```

```
## [1] 18.34703
```

Figura 1. Distribución de edad en la muestra




Fuente: ENDISC−II

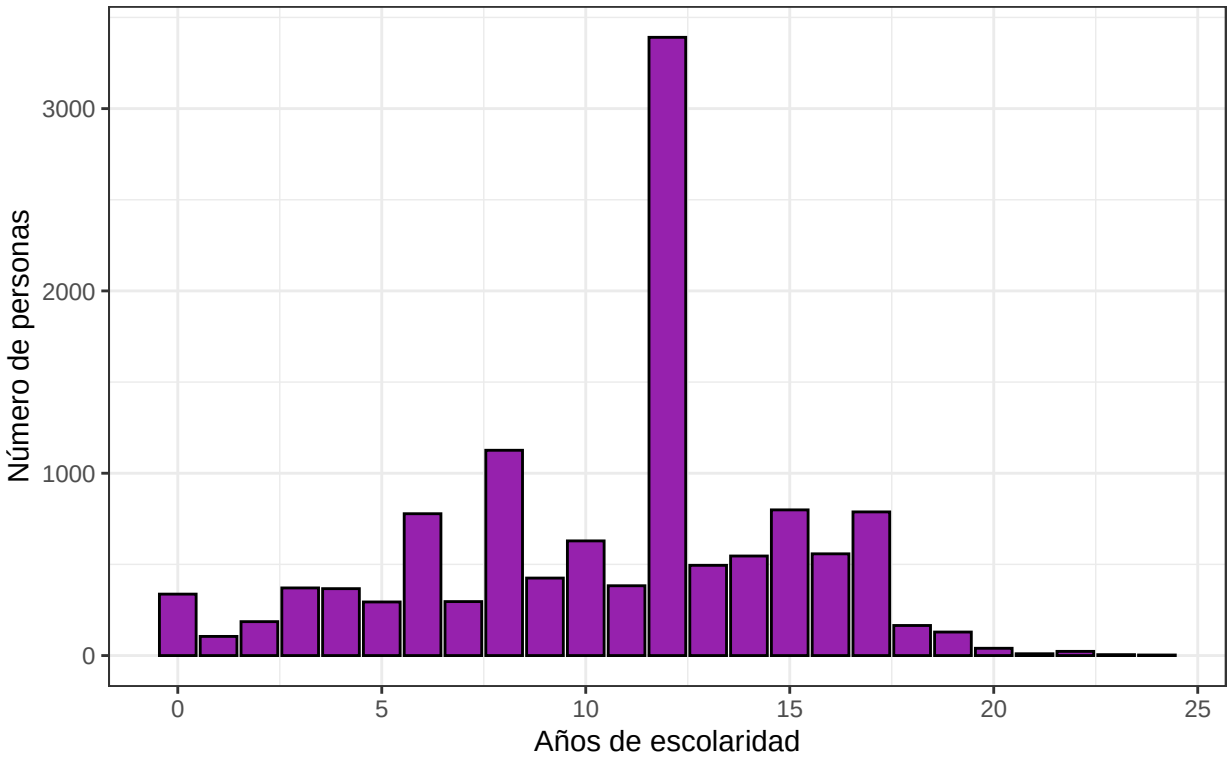
**Educación**

Hay dos variables de educación: grado de escolaridad y años de escolaridad. La primera es una variable categórica ordinal y la segunda es una variable continua cuya distribución no es normal. Podemos usar gráficos de barra y tablas de frecuencia para la primera, y la examinar las medidas de tendencia central e histograma para la segunda.

**Tabla 2. Porcentaje de personas en la muestra por nivel educacional**

| **Nivel educacional** | **%** |
|---|---|
| Sin educacion formal | 2.75 |
| Basica incompleta | 15.56 |
| Basica completa | 11.54 |
| Media incompleta | 14.37 |
| Media completa | 27.84 |
| Superior incompleta | 10.53 |
| Superior completa | 17.41 |

Figura 2. Histograma de años de escolaridad




Fuente: ENDISC−II

```
#Medidas de tendencia central
#Promedio
mean(endisc_cc$esc)
```

```
## [1] 10.67908
```

```
#Mediana
median(endisc_cc$esc)
```

```
## [1] 12
```

**Ingresos**

La variable ingresos es una variable continua. En general, esta variable no tiene una distribución normal. Una gran proporción de población tiene ingresos bajos, mientras que solo una pequeña parte tiene ingresos altos.

```
options(scipen = 50)
#Ingreso promedio y mediano

#Promedio
mean(endisc_cc2$ytot)
```

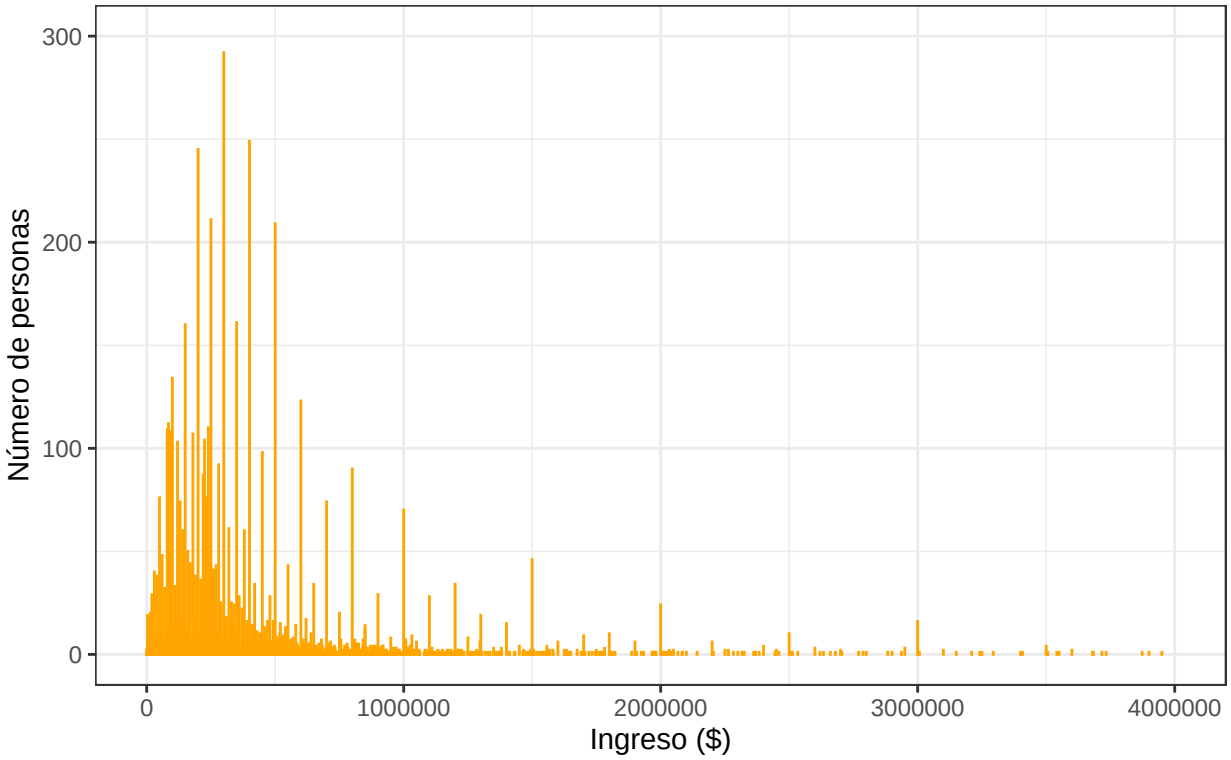
```
## [1] 315785.3
```

```
#Medio
median(endisc_cc2$ytot)
```

```
## [1] 200000
```

```
## Scale for 'x' is already present. Adding another scale for 'x', which will
## replace the existing scale.
```

Figura 3. Distribución de ingresos en la muestra




Fuente: ENDISC−II

**Discapacidad**

Discapacidad es una variable del tipo ordinal. En este caso, voy a utilizar una tabla de frecuencias.

Tabla 3. Número y porcentaje de personas en la muestra por grado de discapacidad

| Grado de discapacidad | N | % |
|---|---|---|
| Sin Discapacidad | 9,633 | 78.64 |
| Leve a Moderada | 1,527 | 12.47 |
| Severa | 1,089 | 8.89 |

**Breakout groups: Utilizando las tablas y figuras realizadas hasta el momento, describe la distribución de las variables de sexo, edad, educación, ingresos y discapacidad en la muestra de ENDISC-II**

## Análisis

Tenemos descritas nuestras variables de interés, ahora analizaremos la relación entre nuestra variable dependiente, discapacidad, con los determinantes sociales de sexo, edad, educación e ingresos.

### Sexo y discapacidad

Tipos de variables? Categórica ordinal y binaria. Puedo utilizar más de una opción

Figura 4. Comparación de grado de discapacidad según sexo

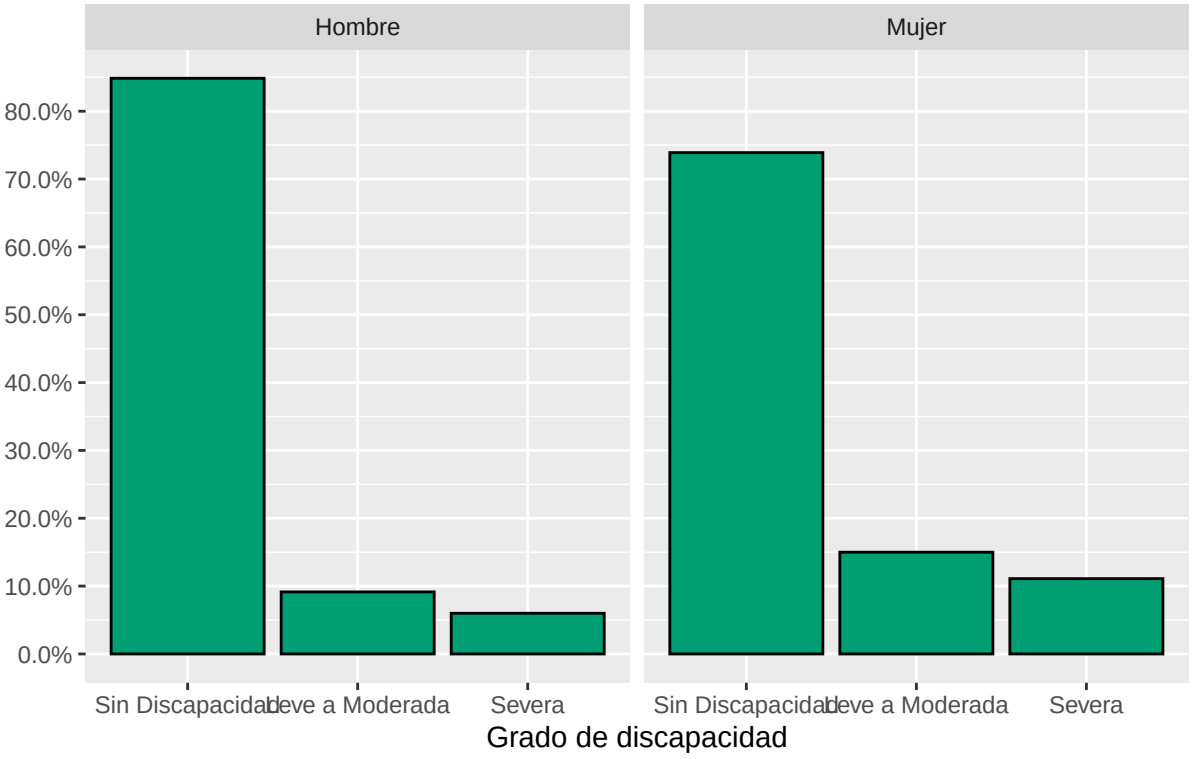
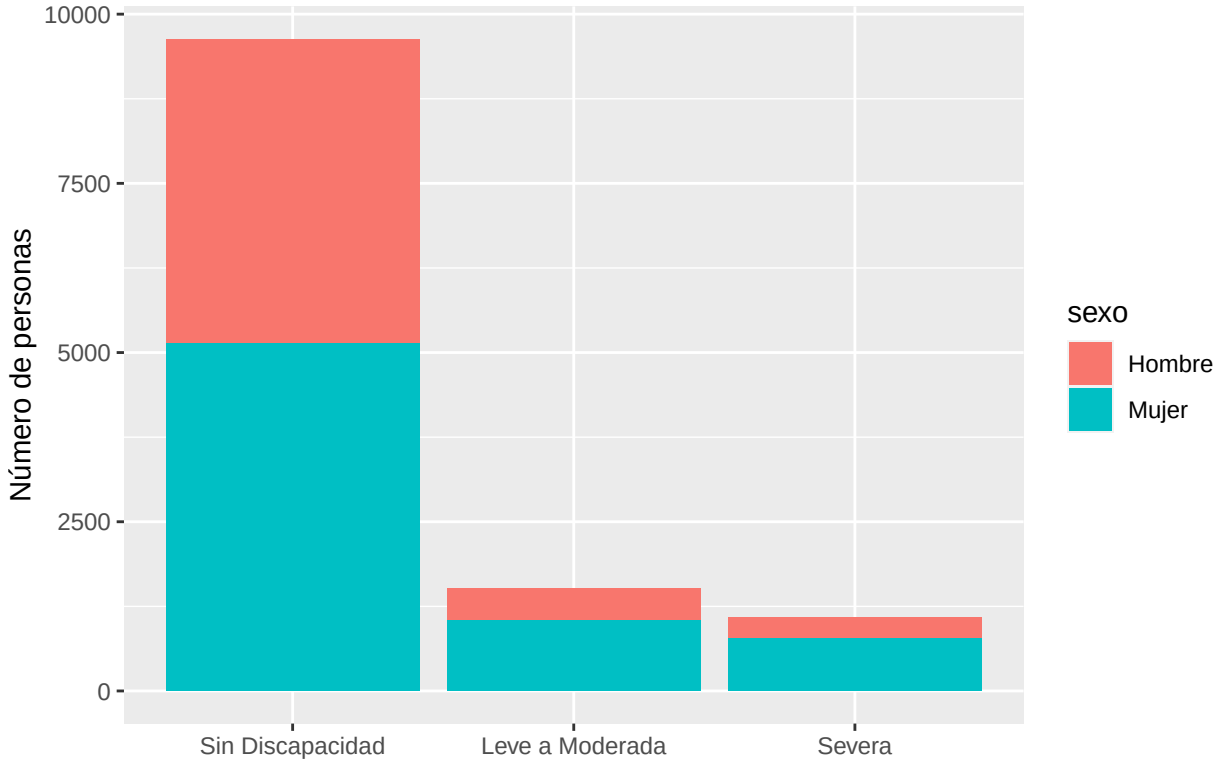


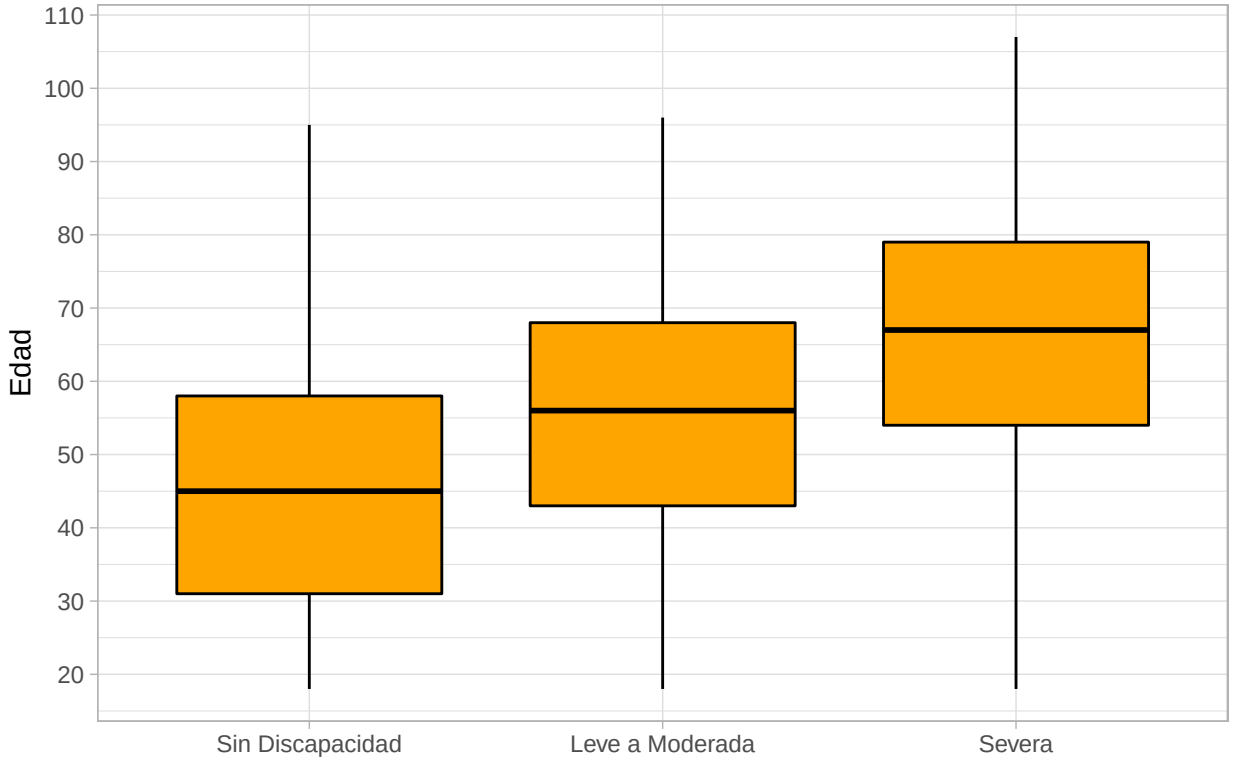
Figura 5. Distribución de discapacidad según sexo




**Edad y discapacidad**

Tipo de variable? Continua y ordinal. Puedo utilizar boxplots

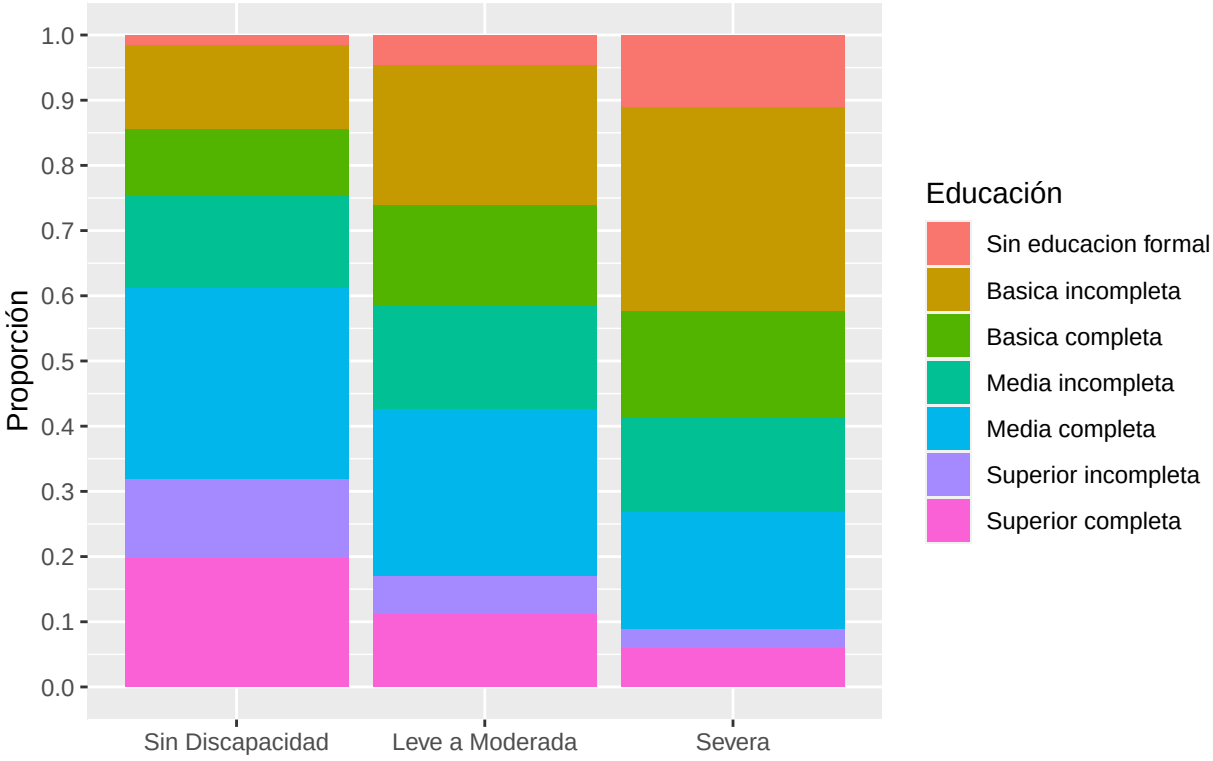
Figura 6. Distribución de discapacidad por edad




**Educación y discapacidad**

Tipo de variable? Grado educativo es ordinal y discapacidad también.

Figura 7. Distribución de educación según discapacidad




**Ingresos y discapacidad**

Ingresos es una variable continua y discapacidad es ordinal. Utilizaré promedio y mediana, sus medidas de tendencia central, para enfatizar las diferencias en una tabla.

**Tabla 5. Mediana y promedio de ingresos por grado de discapacidad**

| Grado de discapacidad | Mediana $ | Promedio $ |
|---|---|---|
| Sin Discapacidad | 228,436 | 345,411 |
| Leve a Moderada | 153,000 | 234,519 |
| Severa | 120,000 | 167,705 |

**Break out groups 2: Utilizando los gráficos y tablas de la sección análisis, describe las asociaciones entre sexo y discapacidad, edad y discapacidad, educación y discapacidad e ingresos y discapacidad.**

**Todo junto**

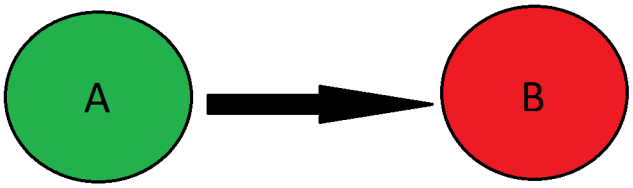
```
unique(endisc_cc %>%
  group_by(disc_grado_adulto) %>%
  mutate(prom_sexo = round(
    200 - mean(as.numeric(sexo))*100,2)) %>%
  mutate(prom_edad = round(mean(edad),2)) %>%
  mutate(prom_educ = round(mean(esc),2)) %>%
    mutate(media_ingr = median(ytot)) %>%
  select(disc_grado_adulto, prom_sexo,
         prom_edad, prom_educ, media_ingr)) %>%
  arrange(disc_grado_adulto) %>% #Crear datos con promedios
  flextable() %>% #Generar tabla en base a los datos
  set_header_labels(disc_grado_adulto =
                      "Grado de discapacidad",
                    prom_sexo = "Porcentaje hombres",
                    prom_edad = "Edad promedio",
                    prom_educ = "Escolaridad promedio",
                    media_ingr = "Ingreso mediano") %>%
  add_header_lines(values = "Tabla 6.Tabla resumen") %>%
  width(j = NULL, width = 1) %>%
  align(align = "left", part = "all") %>%
  theme_vanilla()
```

**Tabla 6.Tabla resumen**

| Grado de discapacidad | Porcentaje hombres | Edad promedio | Escolaridad promedio | Ingreso mediano |
|---|---|---|---|---|
| Sin Discapacidad | 46.71 | 45.33 | 11.30 | 228,436 |
| Leve a Moderada | 31.76 | 55.12 | 9.27 | 153,000 |
| Severa | 29.20 | 65.61 | 7.12 | 120,000 |

## Opcional

Hasta ahora, hemos analizado asociaciones considerando dos variables. Por ejemplo, edad y discapacidad o educación y discapacida.

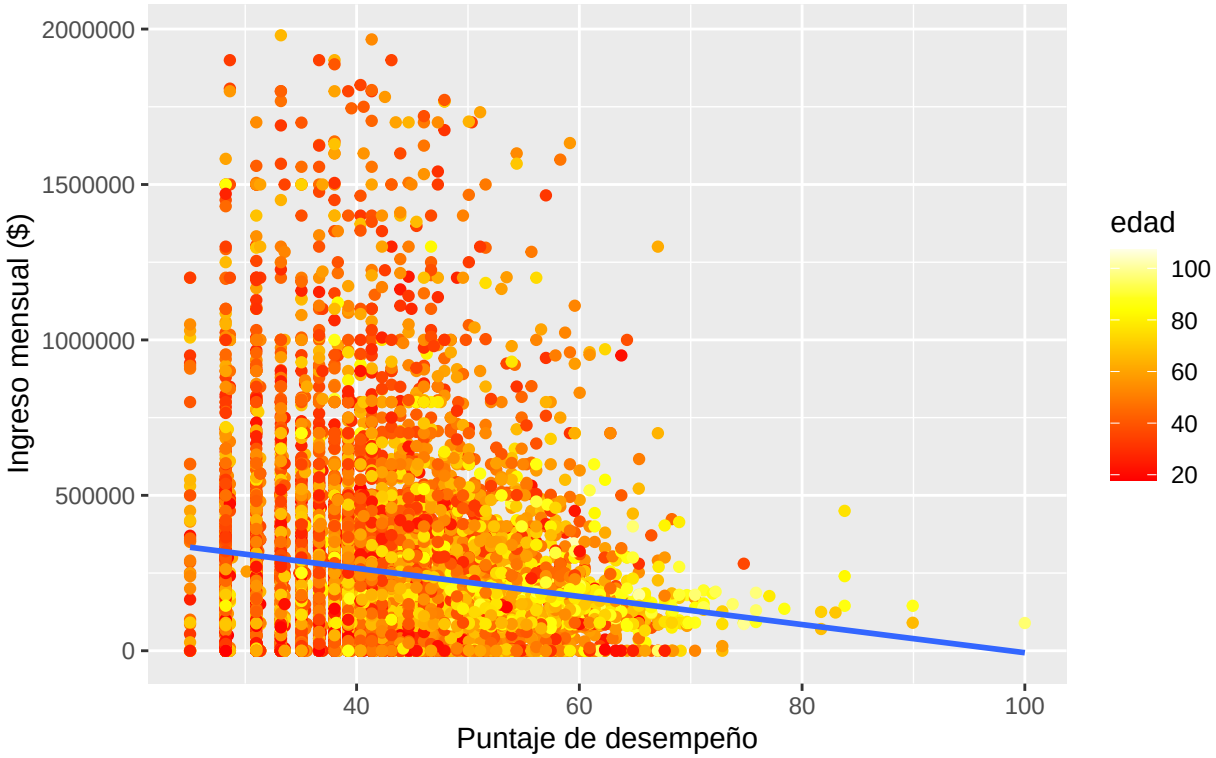



En realidad, van a existir múltiples relaciones actuando de manera simultánea y muchas veces habrán interacciones entre las distintas variables independientes.


Fuente: Lund et al. (2018)

Visualicemos la relación entre tres variables a la vez: ingresos, edad y discapacidad. Cómo es la relación entre estas tres variables?

Figura 8. Scatterplot de la asociación entre puntaje de desempeño e ingreso mensual




Qué podemos decir de la relación que existe entre desempeño, ingresos y edad mirando el gráfico?