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, educación e ingresos; 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.

Notemos los 'NAs' que aparecen. Esto es lo que se llama como missing data. Personas que no tienen información en ciertas variables.

```
##      enc_id          hogar           rph_id          region
##  Min.   :  623   Min.   :1.000   Min.   : 1530   Min.   : 1.000
##  1st Qu.: 4502   1st Qu.:1.000   1st Qu.:11910   1st Qu.: 5.000
##  Median : 9268   Median :1.000   Median :22348   Median : 9.000
##  Mean   : 9074   Mean   :1.024   Mean   :22304   Mean   : 9.013
##  3rd Qu.:13959   3rd Qu.:1.000   3rd Qu.:32645   3rd Qu.:13.000
##  Max.   :17686   Max.   :4.000   Max.   :43180   Max.   :15.000
##
##       sexo            edad            educc             esc
##  Min.   :1.000   Min.   :  0.00   Min.   : 0.000   Min.   : 0.00
##  1st Qu.:1.000   1st Qu.: 18.00   1st Qu.: 1.000   1st Qu.: 8.00
##  Median :2.000   Median : 35.00   Median : 3.000   Median :12.00
##  Mean   :1.524   Mean   : 36.95   Mean   : 3.081   Mean   :10.85
##  3rd Qu.:2.000   3rd Qu.: 54.00   3rd Qu.: 4.000   3rd Qu.:13.00
##  Max.   :2.000   Max.   :108.00   Max.   :99.000   Max.   :24.00
##                                                    NA's   :7917
##       ytot             cie_suma      cap_puntaje_adulto des_puntaje_adulto
##  Min.   :       0   Min.   : 0.000   Min.   :  0.00     Min.   :  0.00
##  1st Qu.:       0   1st Qu.: 0.000   1st Qu.: 14.55     1st Qu.: 24.65
##  Median :   90555   Median : 1.000   Median : 29.46     Median : 38.02
##  Mean   :  220206   Mean   : 1.852   Mean   : 27.95     Mean   : 34.98
##  3rd Qu.:  300000   3rd Qu.: 3.000   3rd Qu.: 39.58     3rd Qu.: 46.68
##  Max.   :26500000   Max.   :12.000   Max.   :100.00     Max.   :100.00
##  NA's   :75         NA's   :22106    NA's   :27621      NA's   :27621
##  disc_grado_adulto
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.302
##  3rd Qu.:0.000
##  Max.   :2.000
##  NA's   :27621

##    Mode   FALSE    TRUE
## logical   12265   27621
```

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)
```

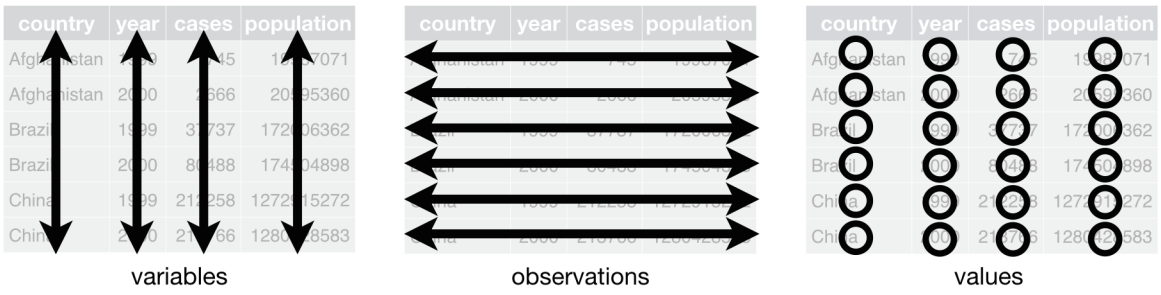
```
## [1] 48.35907
```

```
#Diferencias en nuestra estimación. Necesitamos tener un número determinado que sea constante para todo
#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
```

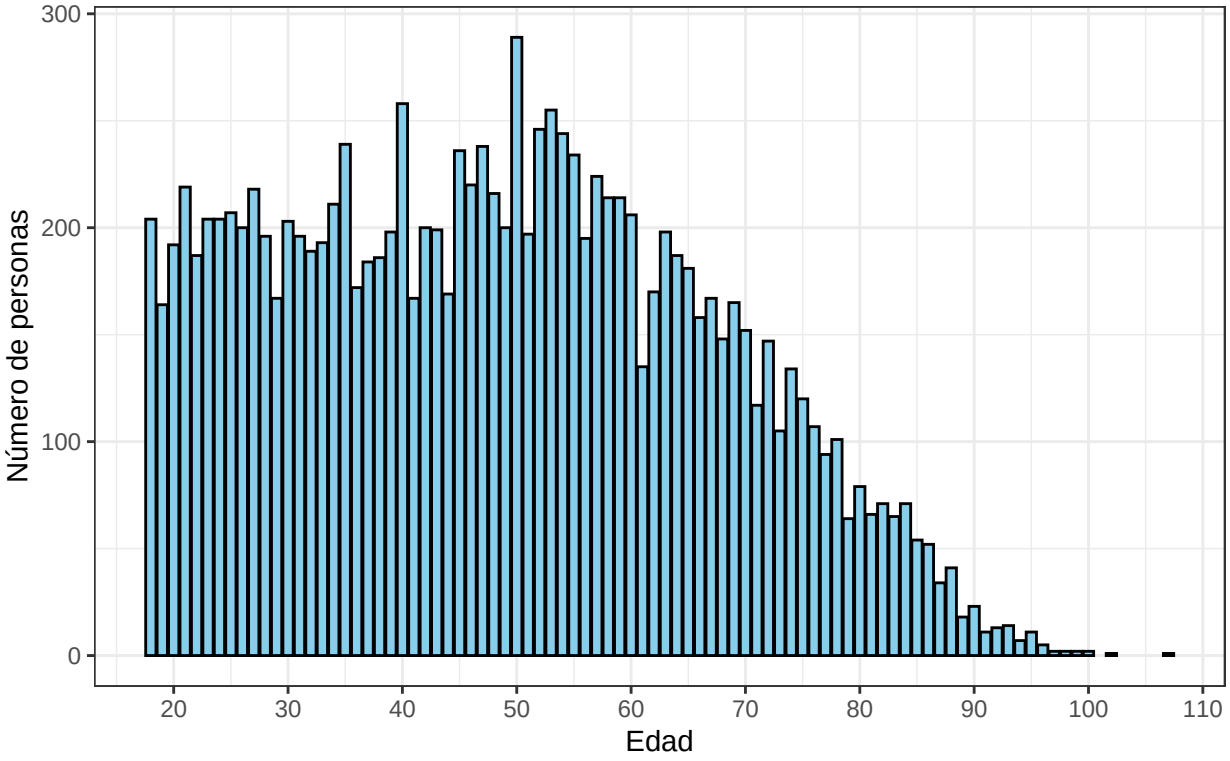
```
#Mediana
median(endisc_cc$edad, na.rm = T)
```

```
## [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

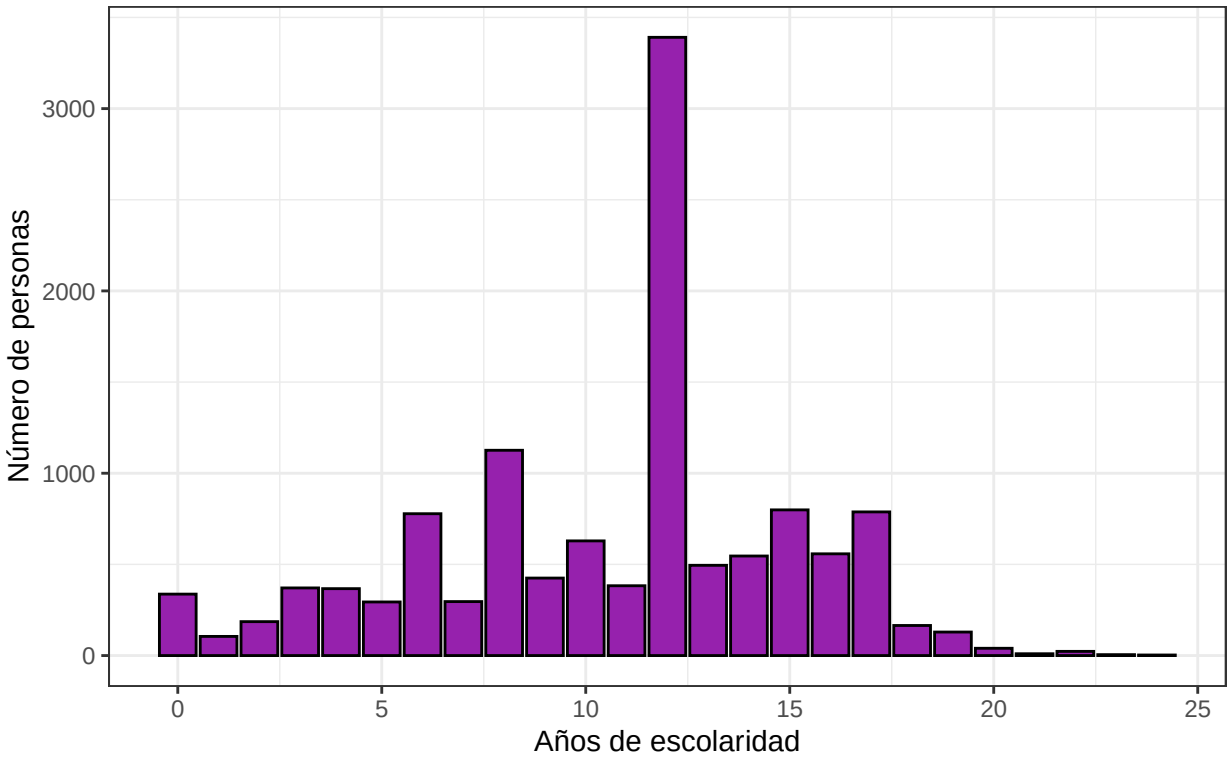
Educación

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)
```

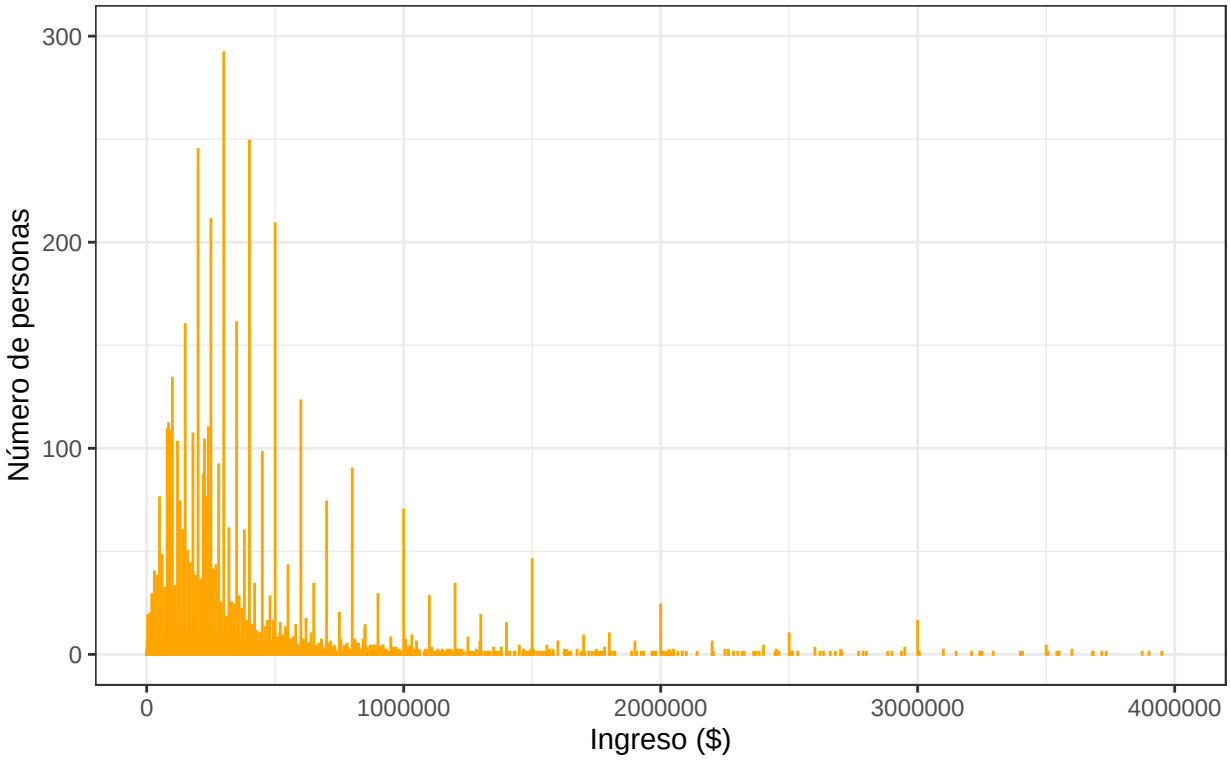
```
## [1] 315785.3
```

```
#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

Sexo: Existe una mayor proporción de mujeres que de hombres en esta muestra. Hay alrededor de 13% más de mujeres que hombres en la muestra.

Edad: El promedio de edad de nuestra muestra es de 48.3 años, cercano a la mediana (48 años). Aquello es un buen indicador acerca de una distribución normal de edad en nuestra muestra. Existe alrededor de 200 personas con la menor edad posible -18 años-, y una cantidad muy pequeña de nuestra muestra es mayor a 90 años. La gran parte de las personas se encuentra entre 18 a 60 años, desde esa edad comienza a disminuir fuertemente el número de personas en la muestra.

Educación: El grupo más grande son las personas con educación media completa, con un 27.84%. Alrededor del 56% de la muestra tiene al menos educación media completa. Esto también se ve reflejado en el histograma, donde el mayor número de personas tiene 12 años de educación. En Chile, desde el año 2003 la educación media se hizo obligatoria. Es posible que aquellos sin educación media completa sean adultos mayores, pero esto es solo una hipótesis ya que no hemos generado insumos para investigar aquello.

Ingresos: Los ingresos en esta muestra no tienen una distribución normal. La gran parte de los ingresos se encuentran concentrados en la parte izquierda del gráfico. Es decir, la mayoría de la muestra tiene ingresos menores a 1 millón de pesos. Solo una pequeña proporción tiene ingresos mayores a 1M. El 50% de la muestra gana $200.000 o menos; mientras que el promedio es de alrededor de 315.000. Esta larga diferencia entre mediana y promedio se da por la distribución no-normal de los datos y la influencia de aquellos con mayores ingresos. En este caso, si tuviésemos que elegir una medida más representativa, sería la mediana.

Discapacidad: La mayor parte de la población no presenta discapacidad, un 80%. Asimismo, existe una gradualidad en el grado de discapacidad. Un 12% posee discapacidad leve a moderada, mientras que un 8% posee una discapacidad severa.

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

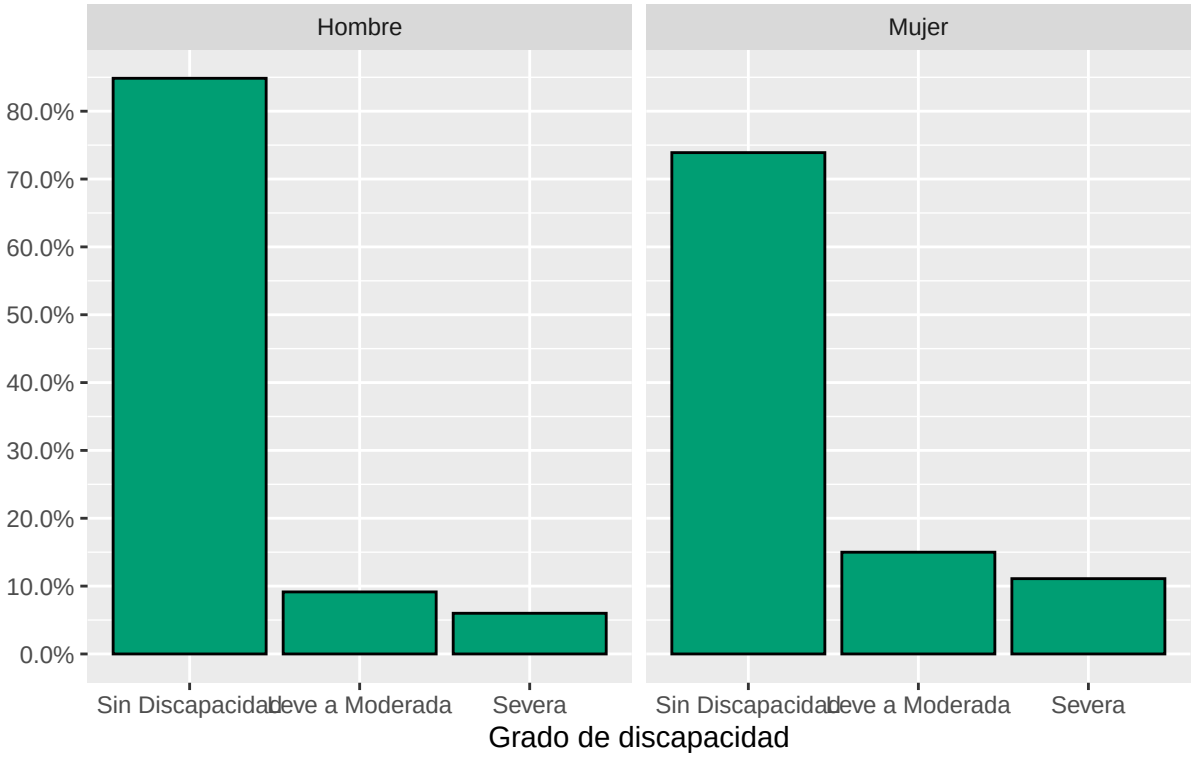
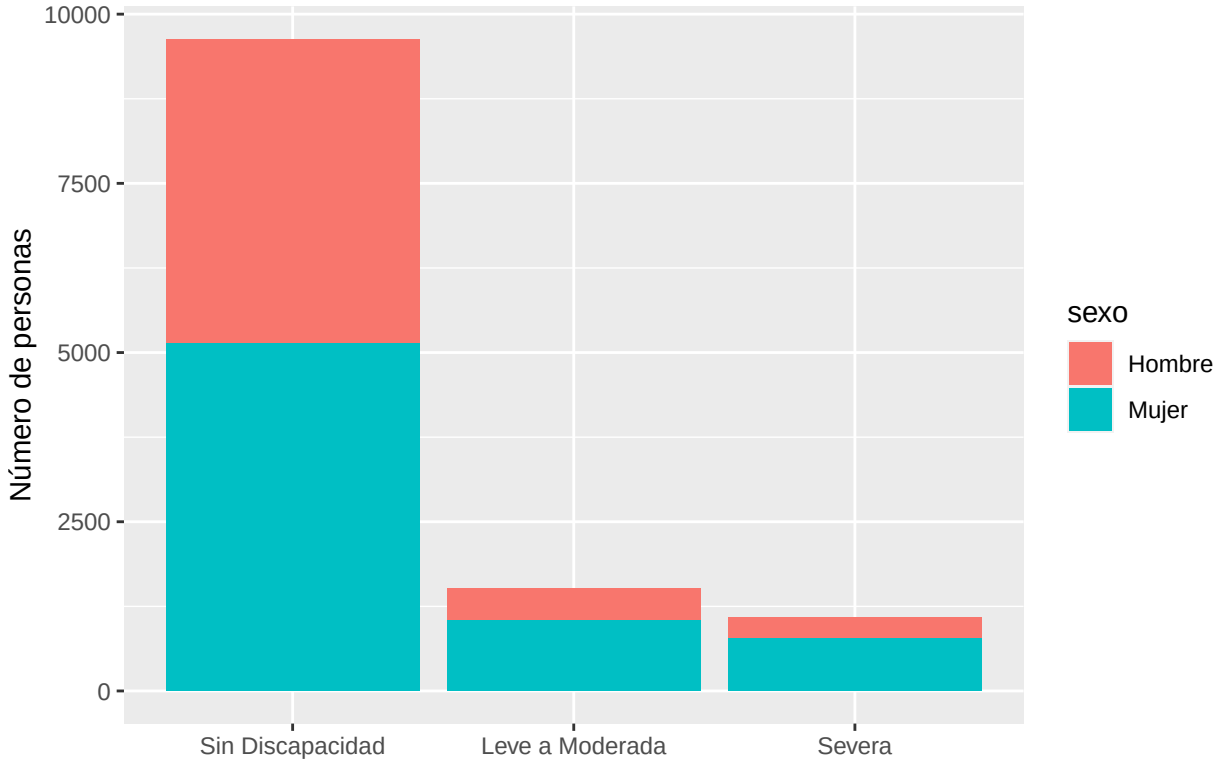


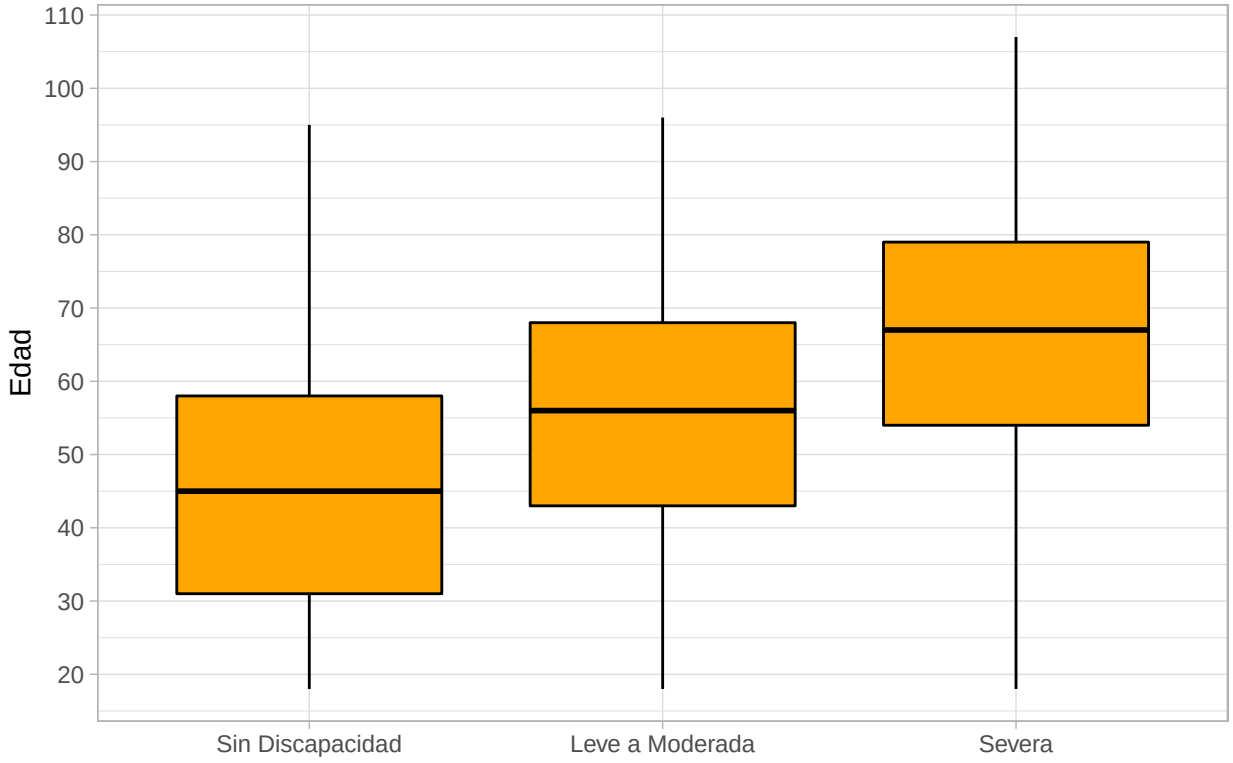
Figura 5. Distribución de discapacidad según sexo



Edad y discapacidad

Tipo de variable? Continua y ordinal. Puedo utilizar boxplots

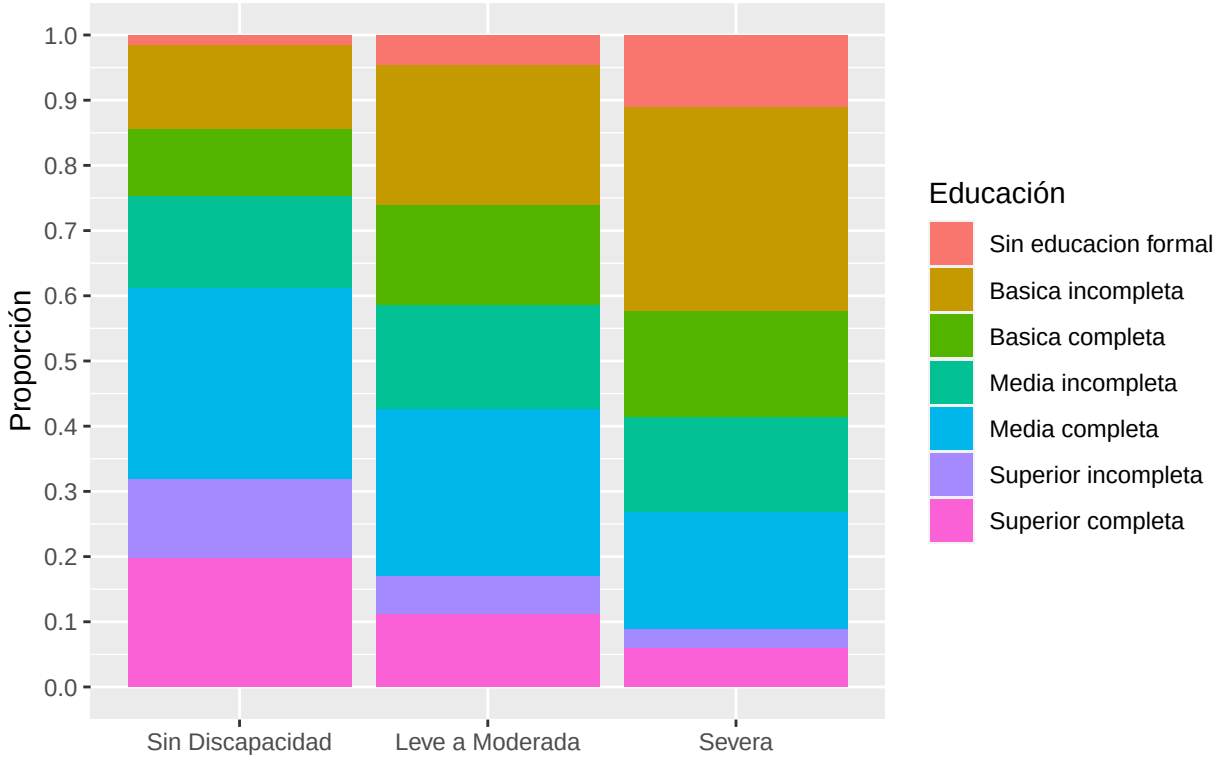
Figura 6. Distribución de discapacidad por edad



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.

Sexo y discapacidad: La Figura 4 y 5 muestran cómo se distribuye el grado de discapacidad en relación a sexo. Ambas figuras apuntan hacia la misma dirección: las mujeres tienen una mayor carga de discapacidad en relación a los hombres. En la Figura 4, Alrededor del 5% de los hombres tienen discapacidad severa, mientras que en el caso de las mujeres, este porcentaje es casi el doble. Algo similar se puede observar en la categoría de discapacidad leve a moderado. En la Figura 5, también se pueden sacar conclusiones similares, sin embargo, es importante recalcar que un gráfico es en base a porcentajes y otro en base a números de

personas. Debido a que existe un mayor número de mujeres en nuestra muestra, es más apropiado utilizar la Figura 4.

Edad y discapacidad: La Figura 6 muestra un gráfico tipo boxplot con la relación entre edad y discapacidad. La línea negra que cruza cada una de las 3 'boxes' representa la media de edad según grado de discapacidad. El cuadro naranjo representa el rango intercuartílico, es decir que el 75% de los datos por grado están contenidos dentro de cada cuadro naranjo. Por último, las líneas negras representan el resto de los datos, en donde se puede observar la edad mínima y máxima según grado de discapacidad. Esta figura sugiere que a mayor edad existe las personas tienden a presentar mayores grados de discapacidad. Asimismo, existe una gradiente según las categorías de discapacidad. La media de edad para aquellos sin discapacidad es de alrededor de 45 años, para aquellos con discapacidad leve o moderada es de 55 años, mientras que para aquellos con discapacidad severa es de 66 años.

Educación y discapacidad: La relación entre educación y discapacidad se puede ver en la Figura 7. Esta figura se llama 'stacked bar chart' (gráfico de barras apiladas). Cada barra representa el 100% de las personas en la categoría de sin discapacidad, leve a moderada, y severa. Debido a que existe un número significativo de categorías, es mejor enfocarse en los extremos. Entre aquellos sin discapacidad, alrededor de un 20% tiene educación superior completa; mientras que para aquellos con discapacidad severa es de alrededor de un 5%. Por otra parte, dentro de aquellos con discapacidad severa, alrededor de un 12% no posee ningún tipo de educación formal, mientras que para aquellos sin discapacidad esta categoría solo representa un 1-2%. Los datos sugieren que las personas con discapacidad tienen un menor nivel educativo en la población chilena.

Ingresos y discapacidad: La Tabla 5 muestra el promedio y la media de ingresos económicos según grado de discapacidad. Debido a lo discutido previamente, nos enfocaremos en la media. La tabla sugiere que personas con grados más severos de discapacidad perciben menos ingresos. Aquellos con discapacidad severa perciben cerca de la mitad de los ingresos que aquellos sin discapacidad. La diferencia de ingresos entre las categorías de discapacidad leve a moderada con severa es relativamente pequeña, del orden de los $30.000.

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()
```

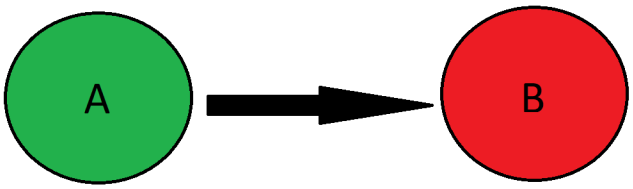
Tabla 6.Tabla resumen

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

Tabla resumen: La Tabla 6 resume los temas discutidos. Solo variables continuas fueron utilizadas para tener una tabla más compacta. En la tabla, se aprecia una menor carga de discapacidad en los hombres -pues solo el 29% de todas las personas con discapacidad severa son hombres-, mayor edad de aquellos con discapacidad, menor escolaridad y también menores ingresos.

Opcional

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

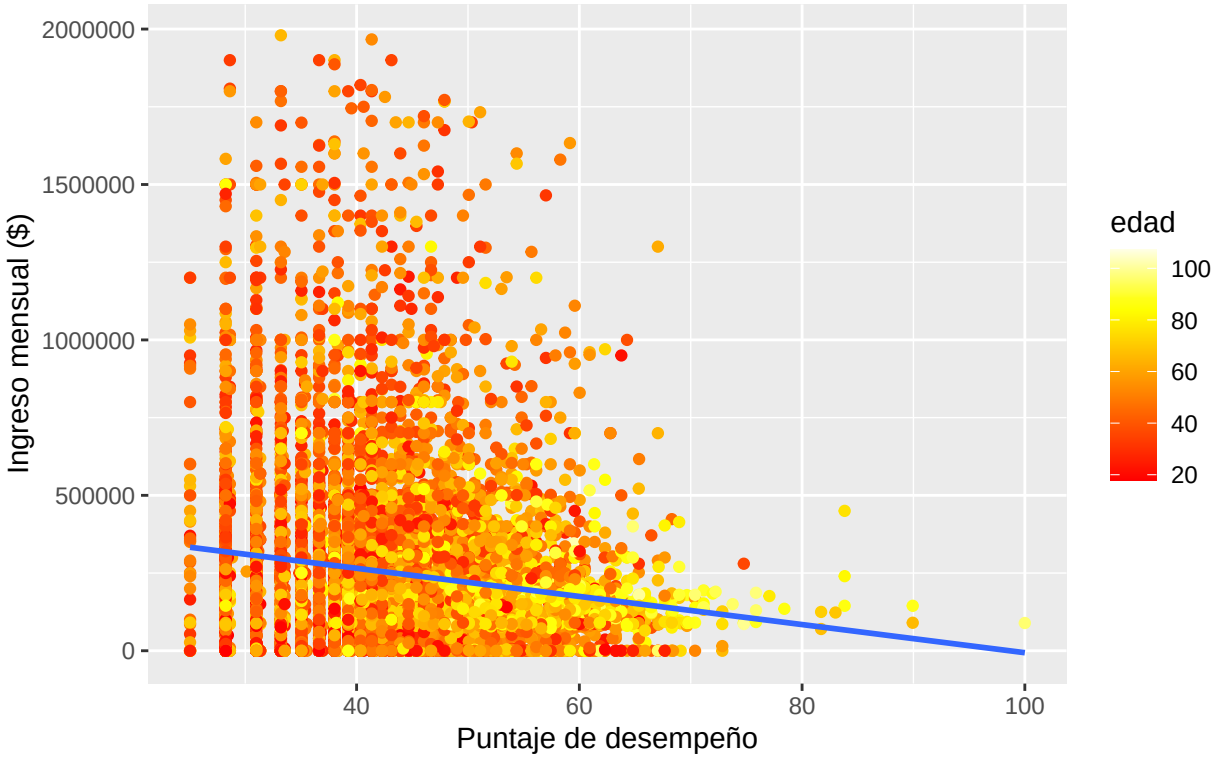


En realidad, van a existir múltiples relaciones actuando de manera simultánea y muchas veces habrán interacciones entre las distintas variables independientes.

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?

Muchas veces nuestra variable de interés está asociada a más de una variable independiente. En el caso de hoy, podemos intuir que el grado de discapacidad está relacionado a factores como el sexo, la edad, la educación y los ingresos. Una de las ventajas de métodos estadísticos más avanzados es que nos permite cuantificar la contribución de estos factores de forma simultánea. A pesar de que estos métodos no son el foco del taller, es bueno tenerlo en cuenta esta simultaneidad en la influencia de la variable dependiente.

En este gráfico, la línea azul representa una regresión linear entre puntaje de desempeño e ingresos. Mayor puntaje en la variable de desempeño significa peor desempeño. Se puede observar una relación indirecta entre ambas variables: a mayor puntaje de desempeño (peor desempeño) las personas tienden a recibir menos ingresos. Sin embargo, este gráfico también nos muestra la edad de las personas. Cada punto es una persona, y el color del punto se relaciona con la edad. Personas más jóvenes están representados con puntos en el espectro más rojo, mientras que personas de mayor edad están representados con puntos en el espectro más amarillo. Si observamos el gráfico, este sugiere que el peor desempeño parece estar asociado a mayor edad, y que menores ingresos también. Es decir, que aquellas personas de peor desempeño suelen estar concentradas entre aquellos más pobres y de mayor edad.