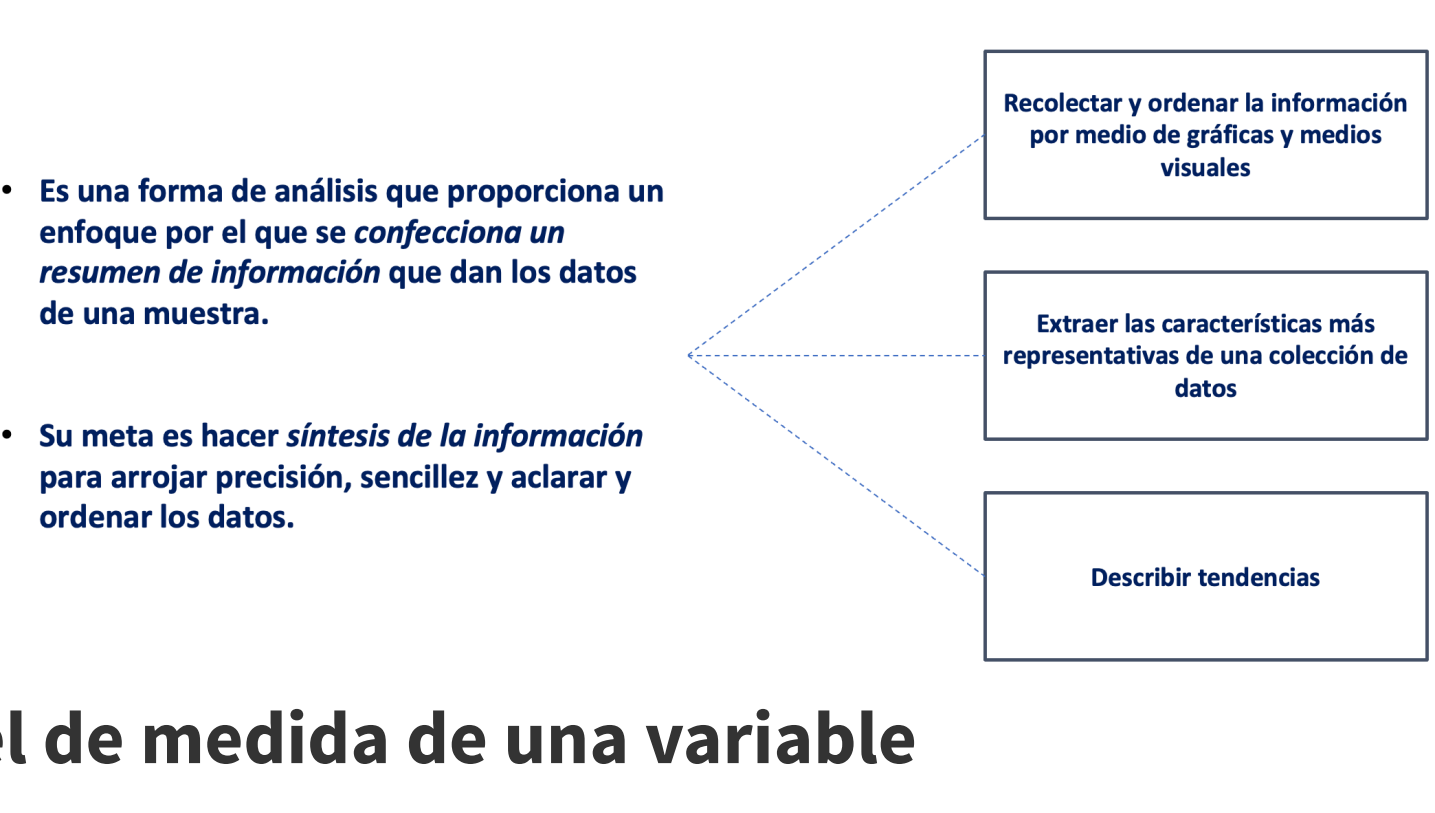
1.¿Qué es el análisis descriptivo?
2.Nivel de medida de una variable
3.Importancia de visualización de datos
4.Análisis descriptivo

Práctica dirigida 2

FACULTAD DE CIENCIAS SOCIALES - PUCP

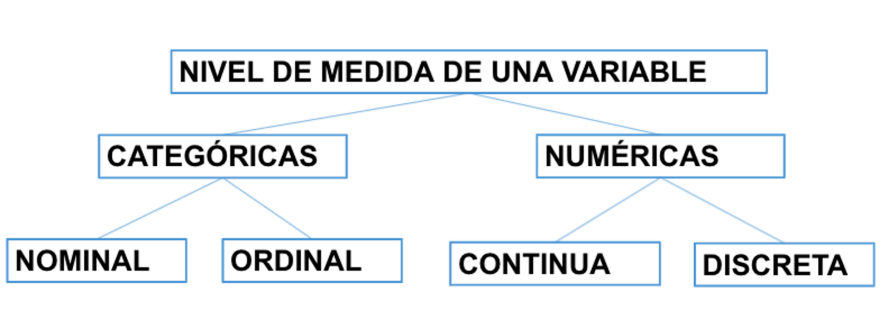
Curso: POL 278 - Estadística para el análisis político 1 | Semestre 2023 - 2

1.¿Qué es el análisis descriptivo?

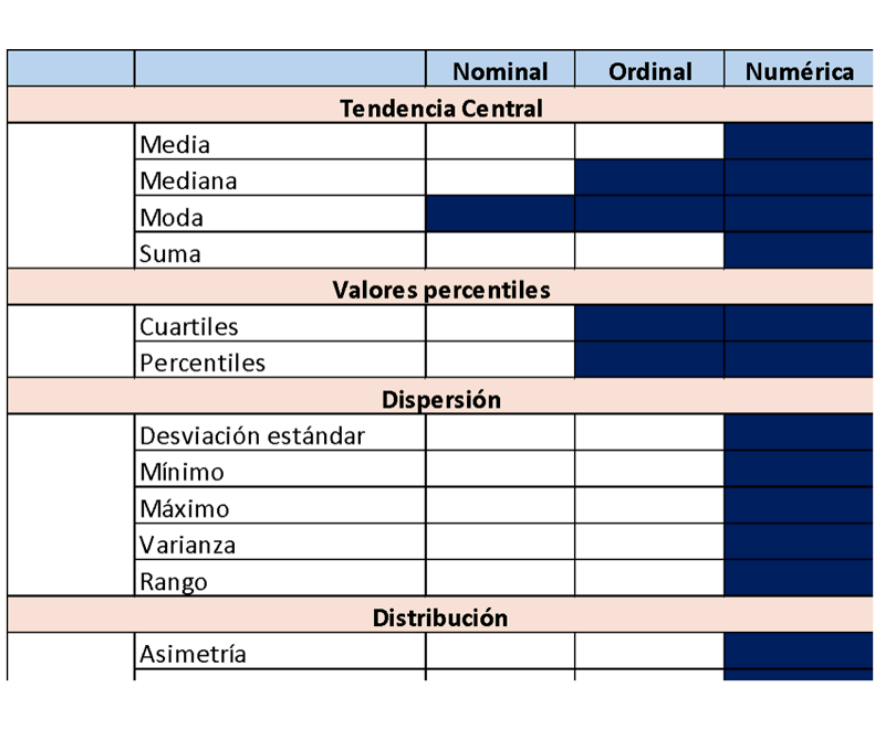


- **Es una forma de análisis que proporciona un enfoque por el que se *confecciona un resumen de información* que dan los datos de una muestra.**
- **Su meta es hacer *síntesis de la información* para arrojar precisión, sencillez y aclarar y ordenar los datos.**

2.Nivel de medida de una variable



Medición por tipo de variables:



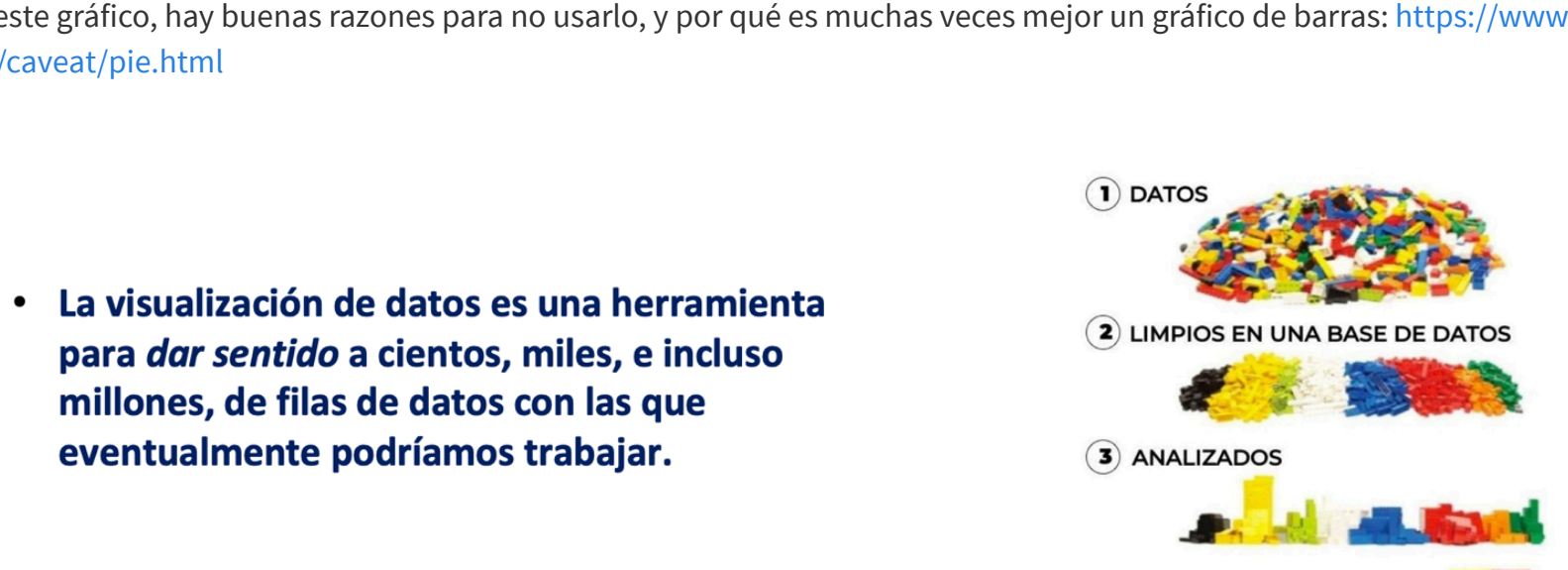
		Nominal	Ordinal	Numérica
Tendencia Central				
	Media			■
	Mediana		■	■
	Moda	■	■	■
	Suma			■
Valores percentiles				
	Cuartiles		■	■
	Percentiles		■	■
Dispersión				
	Desviación estándar			■
	Mínimo			■
	Máximo			■
	Varianza			■
	Rango			■
Distribución				
	Asimetría			■

3.Importancia de visualización de datos

Hay una infinidad de gráficos a los que se puede recurrir dependiendo del interés de la investigadora o el investigador. Aquí hay algunos ejemplos útiles: https://www.data-to-viz.com/

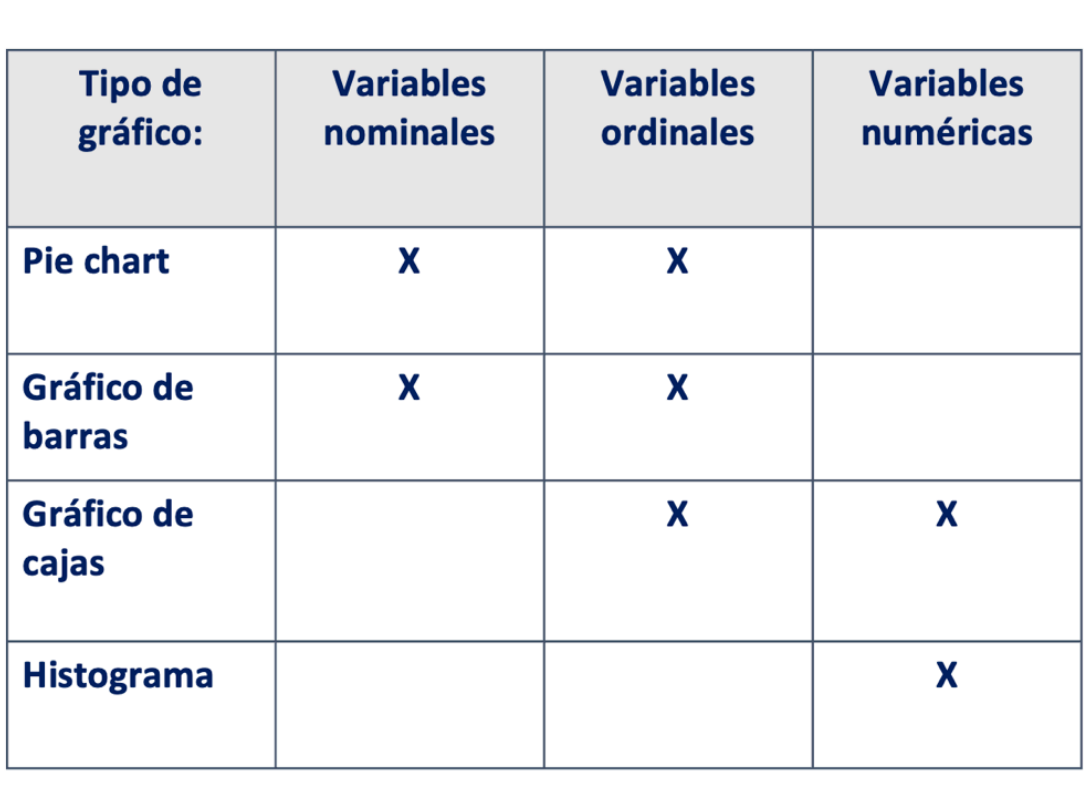
¿Por qué es importante la visualización de datos? Graficar data ayuda a contar historias y, sobre todo, *dar un sentido* a los cientos, miles, o incluso millones, de filas de datos que con las que podríamos eventualmente trabajar, facilitando la comprensión de la información.

En tanto la finalidad de la visualización de datos es ayudar a una mejor comprensión de la información, hay que tener cuidado con algunos gráficos como, por ejemplo, el gráfico de sectores o *pie chart*. A pesar de que la variable que estamos analizando permita utilizar este gráfico, hay buenas razones para no usarlo, y por qué es muchas veces mejor un gráfico de barras: https://www.data-to-viz.com/caveat/pie.html



- **La visualización de datos es una herramienta para *dar sentido* a cientos, miles, e incluso millones, de filas de datos con las que eventualmente podríamos trabajar.**
- **Una adecuada visualización de datos cuenta una historia, eliminando las inconsistencias y "ruidos" en los datos, y resaltando la información útil.**

Gráficos por tipo de variables:

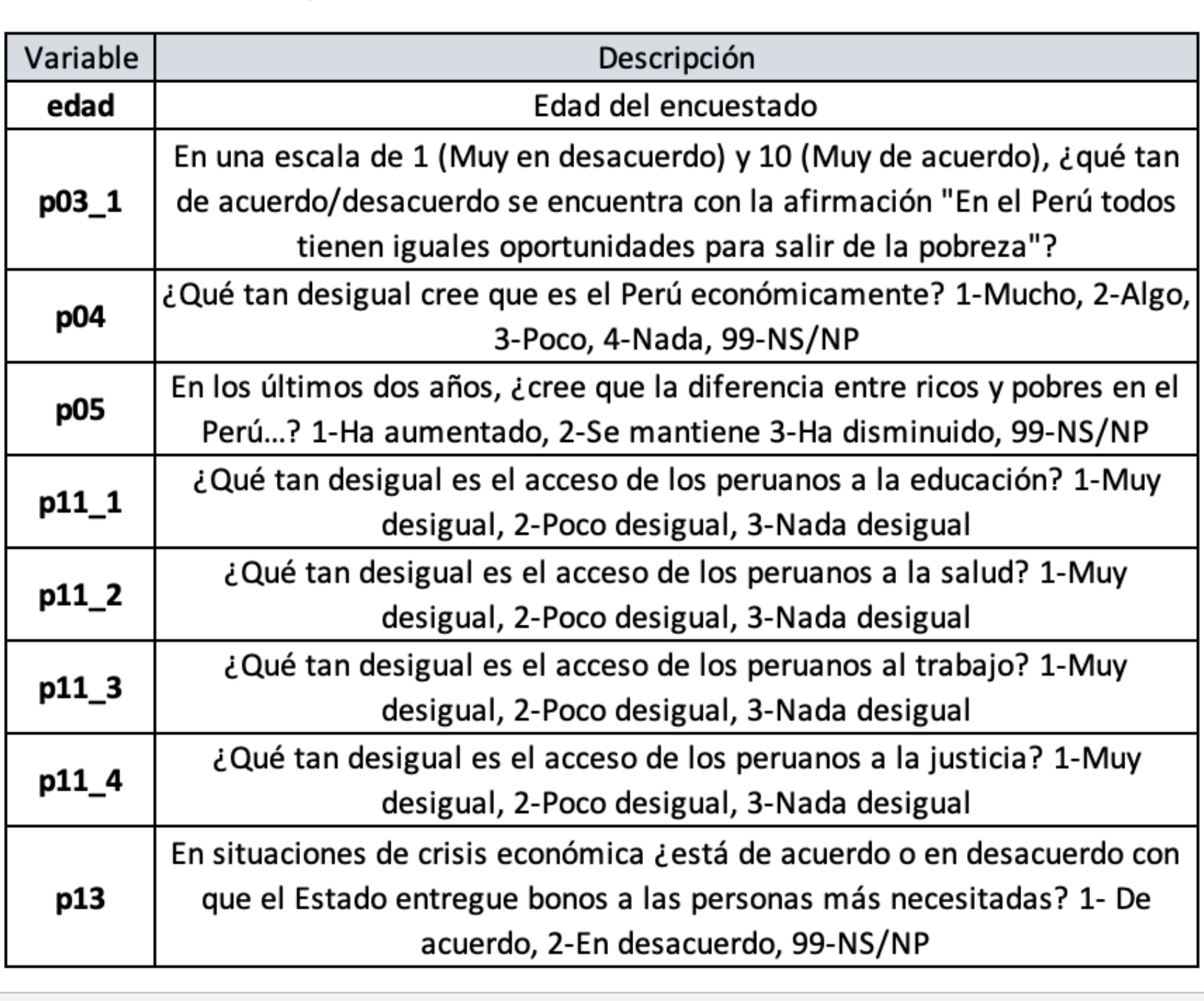


Tipo de gráfico:	Variables nominales	Variables ordinales	Variables numéricas
Pie chart	X	X	
Gráfico de barras	X	X	
Gráfico de cajas		X	X
Histograma			X

4.Análisis descriptivo

Carguemos la data *Enades-2022.dta*.

Hoy trabajaremos con algunas de las variables que forman parte de la Encuesta Nacional de Percepción de Desigualdades - ENADES 2022, que fue elaborada por Instituto de Estudios Peruanos (IEP) y Oxfam. Como lo dice su nombre, esta encuesta busca ahondar en la percepción de las diferentes formas de desigualdad en el Perú e incorpora indicadores que permiten medir la magnitud de brechas sociales y políticas como género, clase, etc.



Variable	Descripción
edad	Edad del encuestado
p03_1	En una escala de 1 (Muy en desacuerdo) y 10 (Muy de acuerdo), ¿qué tan de acuerdo/desacuerdo se encuentra con la afirmación "En el Perú todos tienen iguales oportunidades para salir de la pobreza"?
p04	¿Qué tan desigual cree que es el Perú económicamente? 1-Mucho, 2-Algo, 3-Poco, 4-Nada, 99-NS/NP
p05	En los últimos dos años, ¿cree que la diferencia entre ricos y pobres en el Perú...? 1-Ha aumentado, 2-Se mantiene 3-Ha disminuido, 99-NS/NP
p11_1	¿Qué tan desigual es el acceso de los peruanos a la educación? 1-Muy desigual, 2-Poco desigual, 3-Nada desigual
p11_2	¿Qué tan desigual es el acceso de los peruanos a la salud? 1-Muy desigual, 2-Poco desigual, 3-Nada desigual
p11_3	¿Qué tan desigual es el acceso de los peruanos al trabajo? 1-Muy desigual, 2-Poco desigual, 3-Nada desigual
p11_4	¿Qué tan desigual es el acceso de los peruanos a la justicia? 1-Muy desigual, 2-Poco desigual, 3-Nada desigual
p13	En situaciones de crisis económica ¿está de acuerdo o en desacuerdo con que el Estado entregue bonos a las personas más necesitadas? 1- De acuerdo, 2-En desacuerdo, 99-NS/NP

```
library(rio) #Convocamos el paquete
data=import("Enades_subset.dta")

str(data)
```

```
## 'data.frame':    1390 obs. of  9 variables:
##  $ edad : num  49 60 32 64 19 41 23 19 20 23 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ p03_1: num  5 3 5 10 10 2 1 8 10 1 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ p04  : num  3 1 1 3 1 1 1 2 2 1 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##   ..- attr(*, "labels")= Named num [1:4] 1 2 3 4
##   .. ..- attr(*, "names")= chr [1:4] "Mucho" "Algo" "Poco" "Nada"
##  $ p05  : num  3 1 1 1 1 1 1 1 3 1 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ p11_1: num  2 2 1 2 1 2 1 1 2 1 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ p11_2: num  1 1 1 2 1 1 2 1 2 1 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ p11_3: num  2 1 1 2 1 1 2 1 2 1 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ p11_4: num  1 1 1 2 1 1 2 2 3 1 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ p13  : num  1 2 1 2 1 2 1 2 1 1 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
```

```
names(data) #revisamos los nombres
```

```
## [1] "edad"  "p03_1" "p04"   "p05"   "p11_1" "p11_2" "p11_3" "p11_4" "p13"
```

Análisis de una variable ordinal

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#comprobamos el tipo de dato que analizaremos
class(data$p04)
```

```
## [1] "numeric"
```

Del diccionario de datos, sabemos que esta variable es una ordinal, revisemos los niveles antes de categorizarla correctamente.

```
table(data$p04) #Veamos los niveles de la variable
```

```
##
##   1   2   3   4
## 818 247 250  75
```

Otorguemosle etiquetas y categorizemosla como factor:

```
data$p04 = factor(data$p04, levels = c(1:4), labels = c("Mucho","Algo","Poco","Nada"),ordered =
TRUE)
```

Revisemos que el cambio se haya realizado correctamente correctamente. Para hacerlo, nuevamente Para hacerlo tenemos dos formas, 1. comando table (que revisamos anteriormente), y 2. commando summarize/summarise de dplyr

```
data %>%
 group_by(p04) %>%
  summarize(Freq=n())
```

```
## # A tibble: 4 × 2
##   p04    Freq
##   <ord> <int>
## 1 Mucho   818
## 2 Algo    247
## 3 Poco    250
## 4 Nada     75
```

A primera vista, la tabla nos indica que la mayoría de los encuestados (868) opina que hay mucha desigualdad económica en el país. Pero, ¿cuánto sería eso en porcentaje?

Podemos realizar una tabla de frecuencias y porcentajes agregando una linea al comando anterior

```
para_grafico=data %>%
 group_by(p04) %>%
  summarize(Freq=n()) %>%
   mutate(Porcentaje = (Freq / sum(Freq))*100)
```

Ahora bien, afirmamos que más del 50% de los encuestados percibe que el país es muy desigual económicamente.

Podemos analizar cómo cambia esto si solo seleccionamos los casos de los menores de 30 años.

```
data %>%
  filter(edad<30)%>%
  group_by(p04) %>%
  summarize(Freq=n()) %>%
  mutate(Porcentaje = (Freq / sum(Freq))*100)
```

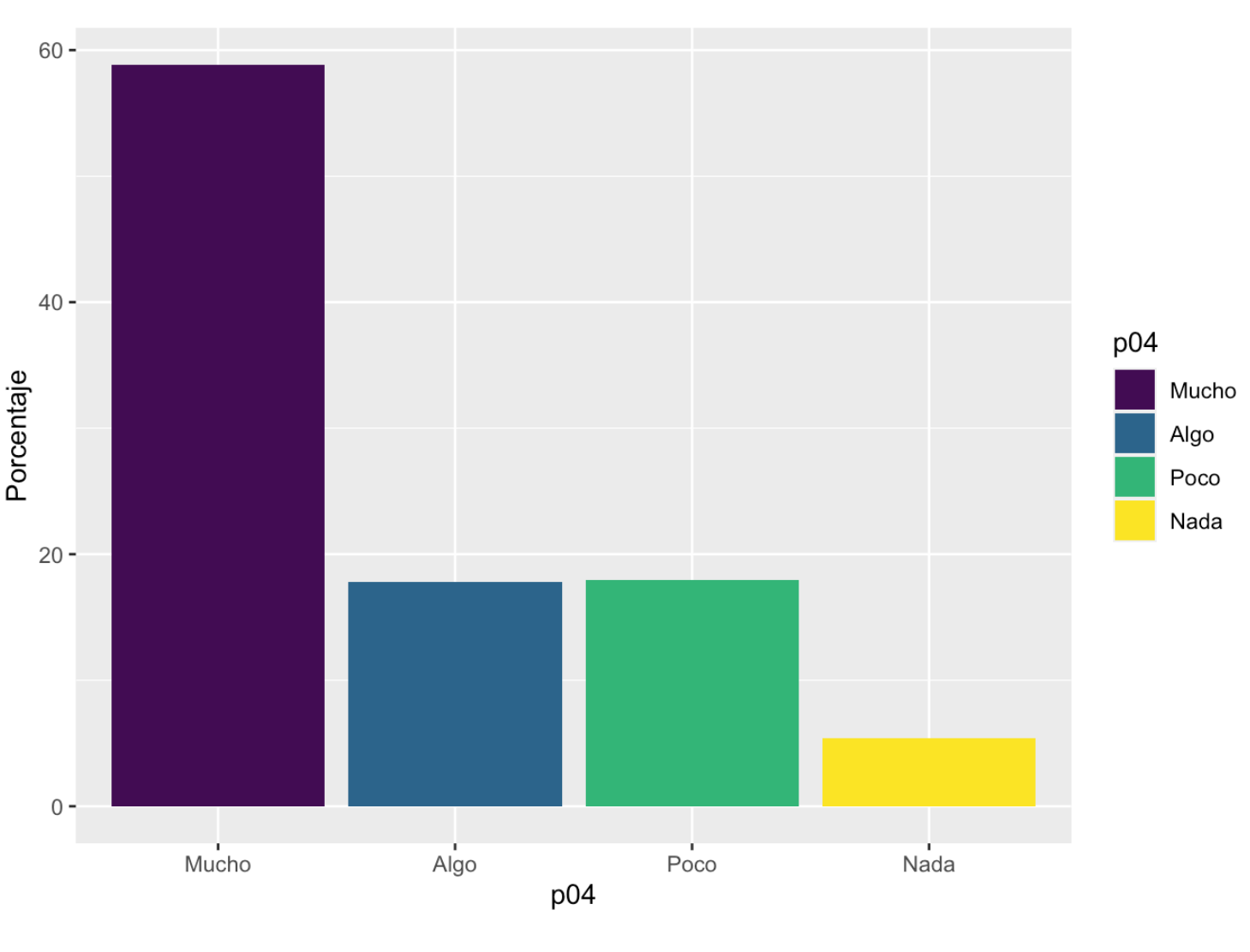
```
## # A tibble: 4 × 3
##   p04    Freq Porcentaje
##   <ord> <int>      <dbl>
## 1 Mucho   264      61.3
## 2 Algo     93      21.6
## 3 Poco     67      15.5
## 4 Nada      7       1.62
```

Grafiquemos los resultados con *ggplot2*

Las tablas creadas anteriormente no fueron guardadas por R como un objeto. Para poder graficar los resultados que arrojó en porcentaje, tendremos que guardarla. Trabajemos con la segunda tabla resumen que contiene los porcentajes y frecuencia sin filtros.

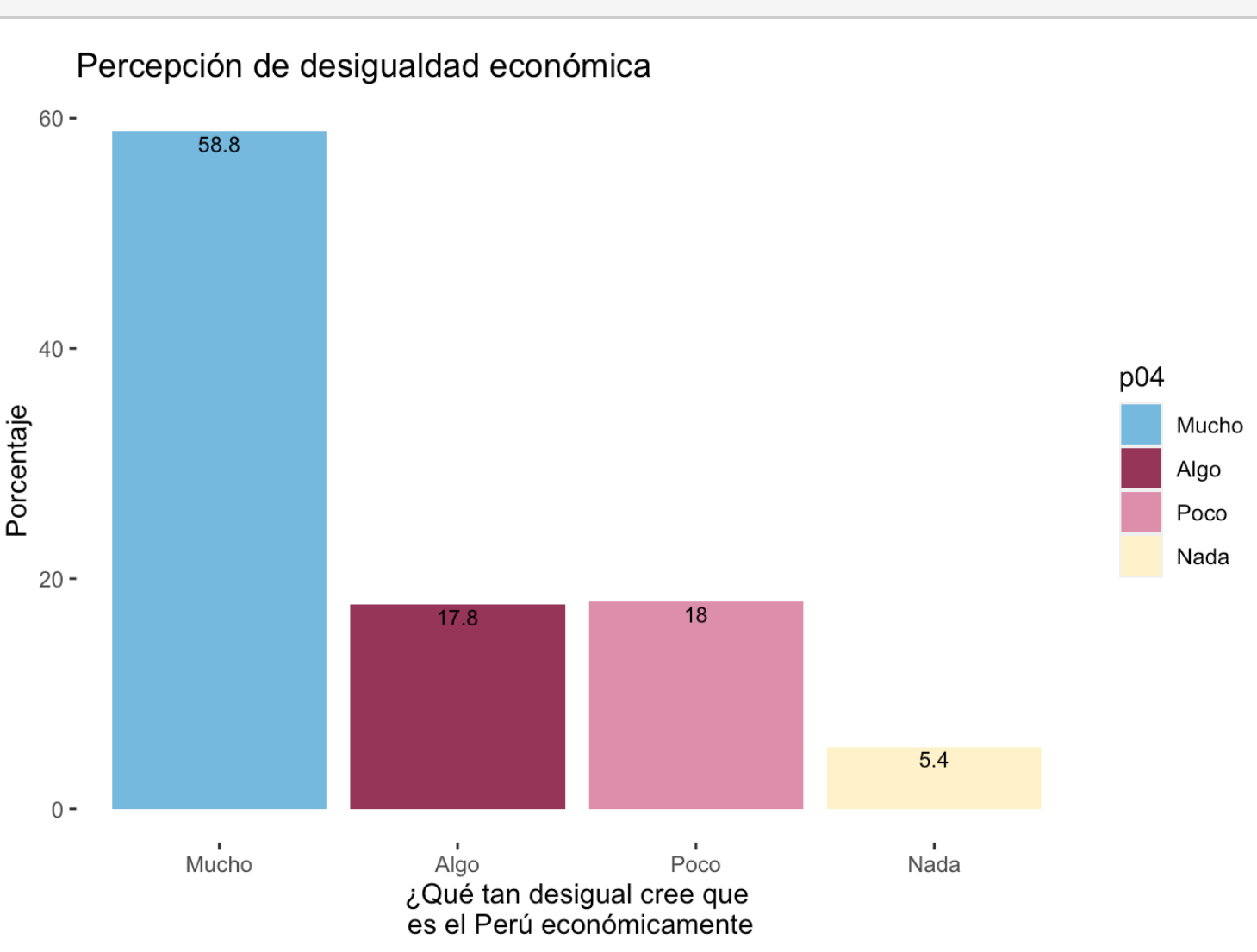
```
library(ggplot2)
library(taylor) #opcional (una ventaja de que R sea software libre)

ggplot(para_grafico, aes(x=p04, y=Porcentaje, fill=p04)) +
geom_bar(stat = "identity")
```



Este es un gráfico básico, pero podemos personalizarlo según nuestros gustos.

```
ggplot(para_grafico, aes(x=p04, y=Porcentaje, fill=p04)) +
  geom_bar(stat = "identity")  +
  ggtitle("Percepción de desigualdad económica") +
  xlab("¿Qué tan desigual cree que\n es el Perú económicamente") + ylab("Porcentaje")+
  geom_text(aes(label=round(Porcentaje,1)), vjust=1.30, color="black", size=3)+
  theme(panel.background=element_rect(fill = "white", colour = "white")) +
  #scale_fill_brewer(palette="Dark2")
  scale_fill_taylor_d(album="Lover") #"Fearless (Taylor's Version)"
```



El ejercicio de análisis descriptivo con variables numéricas lo realizaremos con un indicador aditivo que crearemos a continuación.

Indicador Proxy

También llamado indicador indirecto, se usa ante la imposibilidad de medir lo que efectivamente es de importancia. El indicador mide una variable distinta a la que nos interesa de manera específica, pero presenta una relación lo más directa posible con el fenómeno en estudio.

Un indicador proxy es una medición o señal indirecto que aproxima o representa un fenómeno en la ausencia de una medición o señal directo.

Por ejemplo, el número de miembros femeninos de una cámara de comercio podría ser un indicador proxy para el porcentaje de dueñas de negocios o ejecutivas.

Indicador Aditivo

Pasos para construir un indicador:

1. Verificar que las variables que construyan el indicador correspondan al concepto que se desea medir. *Ejemplo: Si deseo mejor Satisfacción del Usuario, las preguntas deben ser sobre ello.*
2. Revisar el cuestionario e identificar el sentido de las categorías. *Ejemplo: El valor 5 es "Muy instafisfecho" y 1 "Muy satisfecho"*
3. Si las categorías de las variables están en el correcto sentido proceder a sumarlas, si no lo están, proceder a recodificarlas para luego sumar.
4. Una vez realizada la suma, identificar el mínimo y el máximo.
5. Restar a todos los valores el valor mínimo.
6. Al resultado de lo anterior, dividir por el nuevo máximo menos el mínimo, con ello, se va a obtener valores entre 0 y 1.
7. Multiplicar por 100 si se desea el índice de 0 a 100, o por 10 si se desea el índice de 0 a 10.

Construiremos un indicador aditivos de percepción de desigualdad en el acceso a servicios y derechos en el Perú, que vaya del 0 al 100. Para ello usaremos a las variables p11_1,p11_2,p11_3 y p11_4. Estas variables responden a la pregunta de qué tan desigual es el acceso a la eduación, salud, trabajo y justicia; siendo 1-muy desigual y 3- nada desigual.

El indicador que queremos crear es de percepción de desigualdad, por tanto el mayor valor debería ser mayor desigualdad. Para modificarlo podemos cambiar el orden mediante el uso del comando "case_when".

```
data=data %>%
 mutate(d_educ=case_when(
  p11_1 == 1 ~ "3",
  p11_1 == 2 ~ "2",
  p11_1 == 3~ "1"),
  d_salud=case_when(
  p11_2 == 1 ~ "3",
  p11_2 == 2 ~ "2",
  p11_2 == 3~ "1"),
  d_trabajo=case_when(
  p11_3 == 1 ~ "3",
  p11_3 == 2 ~ "2",
  p11_3 == 3~ "1"),
   d_justicia=case_when(
  p11_4 == 1 ~ "3",
  p11_4 == 2 ~ "2",
  p11_4 == 3~ "1"))
```

Revisemos que se haya realizado correctamente

```
table(data$p11_1)
```

```
##
##   1   2   3
## 857 463  70
```

```
table(data$d_educ)
```

```
##
##   1   2   3
##  70 463 857
```

Para poder crear el indicador, necesitamos que todas las variables a usar sean numéricas porque las tendremos que sumar.

```
str(data)#Podemos notar que las variables que creamos son de tipo caracter.
```

```
## 'data.frame':    1390 obs. of  13 variables:
##  $ edad      : num  49 60 32 64 19 41 23 19 20 23 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ p03_1     : num  5 3 5 10 10 2 1 8 10 1 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ p04       : Ord.factor w/ 4 levels "Mucho"<"Algo"<..: 3 1 1 3 1 1 1 2 2 1 ...
##  $ p05       : num  3 1 1 1 1 1 1 1 3 1 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ p11_1     : num  2 2 1 2 1 2 1 1 2 1 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ p11_2     : num  1 1 1 2 1 1 2 1 2 1 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ p11_3     : num  2 1 1 2 1 1 2 1 2 1 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ p11_4     : num  1 1 1 2 1 1 2 2 3 1 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ p13       : num  1 2 1 2 1 2 1 2 1 1 ...
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ d_educ    : chr  "2" "2" "3" "2" ...
##  $ d_salud   : chr  "3" "3" "3" "2" ...
##  $ d_trabajo : chr  "2" "3" "3" "2" ...
##  $ d_justicia: chr  "3" "3" "3" "2" ...
```

Recategorizemos nuestras variables a numéricas:

```
names(data)
```

```
##  [1] "edad"       "p03_1"      "p04"        "p05"        "p11_1"
##  [6] "p11_2"      "p11_3"      "p11_4"      "p13"        "d_educ"
## [11] "d_salud"    "d_trabajo"  "d_justicia"
```

```
data[9:13] = lapply(data[9:13],as.numeric) #podemos usar lapply para recodificar más de una vari
able a la vez
```

Recordemos los pasos para crear un índice aditivo:

```
*((var_suma - mín_de_suma)/(máx_de_suma-mín de suma)))\*valor al que quiere que llegue el índic
e(si va del 0 al 10 será 10, del 0 al 50 será 50, etc).*
```

```
data=data %>%
  mutate(suma = d_educ + d_salud + d_trabajo + d_justicia)
```

Revisamos mínimo y máximo

```
summary(data$suma)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00   10.00   11.00   10.64   12.00   12.00
```

```
data = data %>%
 mutate(indicador = ((suma-4)/(12-4))*100) #Como queremos que el indicador vaya del 0 al 10, lo
multiplicamos por 10
```

Análisis de variables numéricas

Ahora, veamos algunas medidas de tendencia central, distribución y dispersión para el caso de variables numéricas. Trabajaremos con el indicador que acabamos de crear

Exploremos la variable. Veamos medidas de tendencia central y de dispersión. Recordemos que va del 0 al 100.

```
data%>%
  summarise(Media = mean(indicador),
            Mediana = median(indicador),
            Desviacion = sd(indicador),
            Minimo = min(indicador),
            Maximo = max(indicador))
```

```
##     Media Mediana Desviacion Minimo Maximo
## 1 83.0036    87.5   18.63232      0    100
```

Podemos analizar la respuesta según si están de acuerdo con la entrega de bonos (p13 - 1 De acuerdo y 2-Desacuerdo)

```
data$p13 = factor(data$p13, levels = c(1:2), labels = c("De acuerdo","Desacuerdo"),ordered = TRU
E)
```

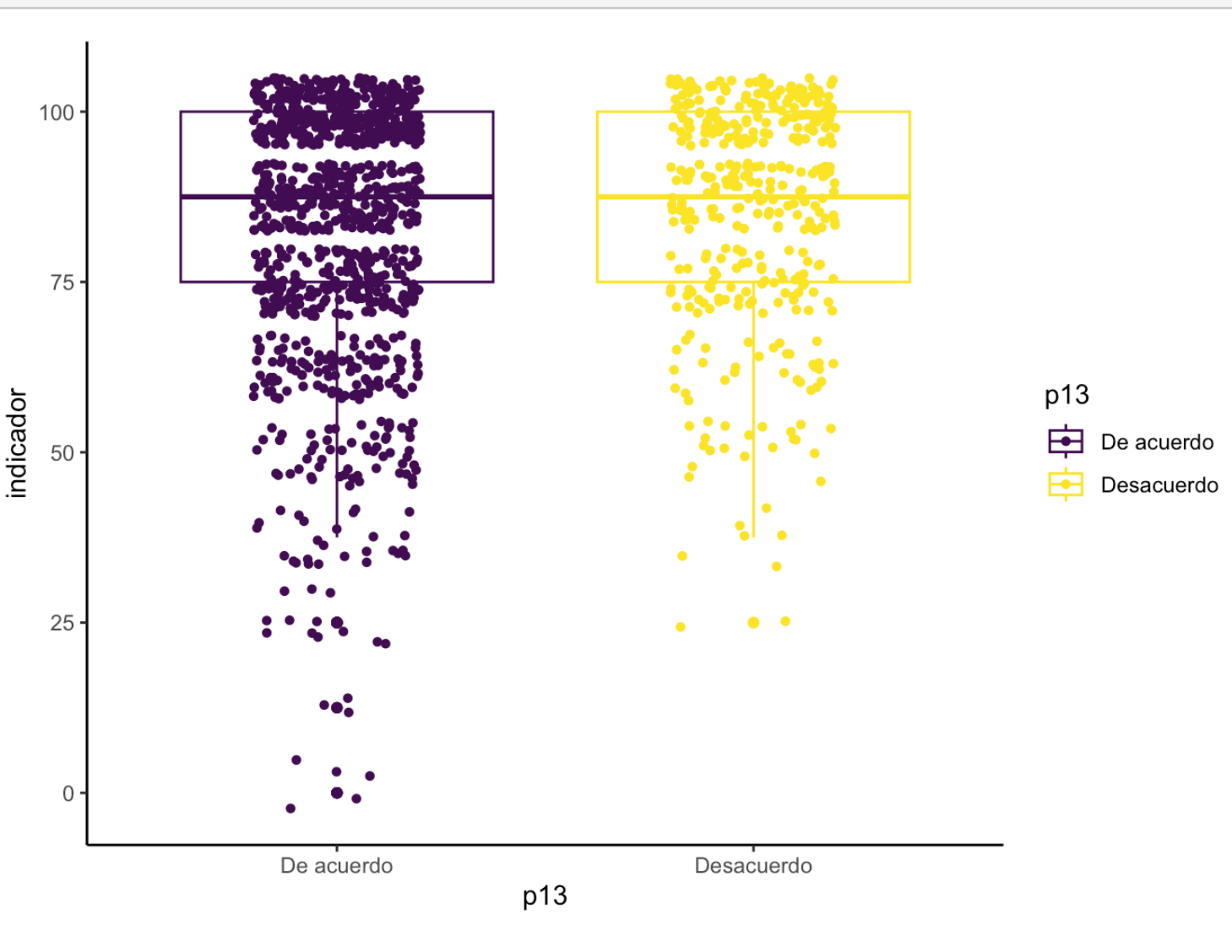
```
data %>%
 group_by(p13) %>%
  summarize(Media=mean(indicador))
```

```
## # A tibble: 2 × 2
##   p13         Media
##   <ord>       <dbl>
## 1 De acuerdo   82.1
## 2 Desacuerdo   85.4
```

La tabla nos indica que aquellos encuestados/as que están de acuerdo con la entrega de bonos, perciben ligeramente una menor desigualdad en los derechos y servicios.

Podemos visualizarlo mejor con un gráfico

```
ggplot(data, aes(x=p13, y=indicador, color=p13)) +
  geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2)) +#para agregar los casos como puntos
  theme_classic()
```



Ejercicio: Analice descriptivos y elabore el gráfico correspondiente para la variable edad.