

Reporte de Base de Datos

Tercer reporte



Grupo : 4

Integrantes : -Abraham Calderón Canicoba(20203599)
-Jean Pierre Arroyo Gonzales(20211199)
-Jimena Solange López Penadillo(20211552)
-Geraldine Namie Pajuelo Sanchez(20193790)

Facultad : Ciencias Sociales

Profesor : José Mendoza

EJERCICIO 1

¿Existe alguna relación entre el nivel educativo de los jefes/as de hogar y el nivel educativo de los miembros más jóvenes del hogar?

Es necesario aclarar que con los miembros más jóvenes nos referimos a los mayores a 25 años. Es decir con esta pregunta queremos saber si el nivel alcanzado por los jefes/as de hogar tiene alguna influencia en el nivel alcanzado por los miembros jóvenes. No obstante para no complicarnos con el análisis, solo dividiremos el nivel alcanzado por los jefes/as de hogar en dos grupos: Educación superior completa y su complemento (no universitaria completa). Para la responder la pregunta los módulos 200 y 300 de ENAHO(Encuesta Nacional de hogares) de los años 2013 al 2022.

Subiendo las base de datos de los 10 años de los 200 y 300 :

```
set more off
clear
global dir "C:\Users\USER\Downloads"

global datadir "C:\Users\USER\Downloads\ENAHOP3"

foreach x of numlist 2013/2022{
    use "$datadir\enaho01-`x'-200", clear

    merge 1:1 conglome vivienda hogar codperso using "$datadir\enaho01-`x'-200", keep(3) nogen

    merge 1:1 conglome vivienda hogar codperso using "$datadir\enaho01a-`x'-300", keep(3) nogen

    gen year = `x'
    destring(codtarea codtiempo), replace
    tempfile b`x'
    save `b`x''
}

use `b2013'

append using `b2014', force
append using `b2015', force
append using `b2016', force
append using `b2017', force
append using `b2018', force
append using `b2019', force
append using `b2020', force
append using `b2021', force
append using `b2022', force
```

Identificando a los jefes de hogar que tienen educación superior universitaria o no universitaria completa :

```
gen jefedusup = .
replace jefedusup = 1 if p203b == 1 & (p301a == 8 | p301a == 10)
replace jefedusup = 0 if p203b == 1 & (p301a != 8 & p301a != 10)
```

Clasificando a todos los hogares de acuerdo a: 1 para todos los miembros si el jefe/a de hogar tiene educación superior completa y 0 en caso contrario :

```
egen grupo = group(conglome vivienda hogar)

gen filtro_hogar = (p203a == 1 & jefedusup == 1)
egen hedusup = max(filtro_hogar), by(grupo)
drop filtro_hogar
```

Construyendo el indicador de años de educación para todos los miembros :

```
gen study_years = p301b

replace study_years= 0 if p301a == 1 | p301a == 2 // Sin nivel o Inicial

recode study_years (1=1) (2=2) (3=3) (4=4) if p301a==3 // Primaria incompleta
recode study_years (5=5) (6=6) if p301a==4 // Primaria completa
recode study_years (1=7) (2=8) (3=9) (4=10) if p301a==5 // Sec. incompleta
recode study_years (4=11) (5=11)(6=12) if p301a==6 // Sec. completa
recode study_years (1=12)(2=13)(3=14)(4=15) if p301a==7 // Sup. No Uni. incompleta
recode study_years (3=14)(4=15)(5=16) if p301a==8 // Sup. No Uni. completa
recode study_years (1=12)(2=13)(3=14)(4=15)(5=16) (6=17) if p301a==9 // Sup. Uni incompleta
recode study_years (4=15)(5=16)(6=17)(7=18) if p301a==10 // Sup. Uni completa
recode study_years (1=17)(2=18) if p301a==11 // Posgrado
```

Obteniendo el promedio de años de educación para mayores de 25 años para cada grupo:
Con este promedio podemos comparar las medias de ambos grupos, es decir el esperado. De esta forma se podrá ver cuanto se esperaría que fuera la diferencia de años de educación de un miembro de cada grupo.

CODIGO:

```
* Para los hogares con jefes de hogar con educación superior completa:
mean(study_years) if p208a > 25 & jefedusup == 1

*Para los hogares con jefes de hogar que no cuentan con educación superior completa:
mean(study_years) if p208a > 25 & jefedusup == 0
```

RESULTADO:

-Media del grupo con jefes(as) de hogar con superior completar

	Mean	Std. err.	[95% conf. interval]	
study_years	15.25975	.0042411	15.25144	15.26807

-Media del otro grupo

	Mean	Std. err.	[95% conf. interval]	
study_years	6.513632	.0089369	6.496116	6.531148

INTERPRETACIÓN El promedio de años de estudio de los miembros jóvenes de una familia con jefe de hogares es de un poco más de 15 años, mientras que de los miembros jóvenes con jefes de hogares que no culminaron la universidad, es de 6 años y medio, una diferencia considerable. Aparte de ello es necesario destacar que los miembros de familia de familias con jefes que no completaron la universidad es de 317 112, mientras que el otro grupo es 61 304.

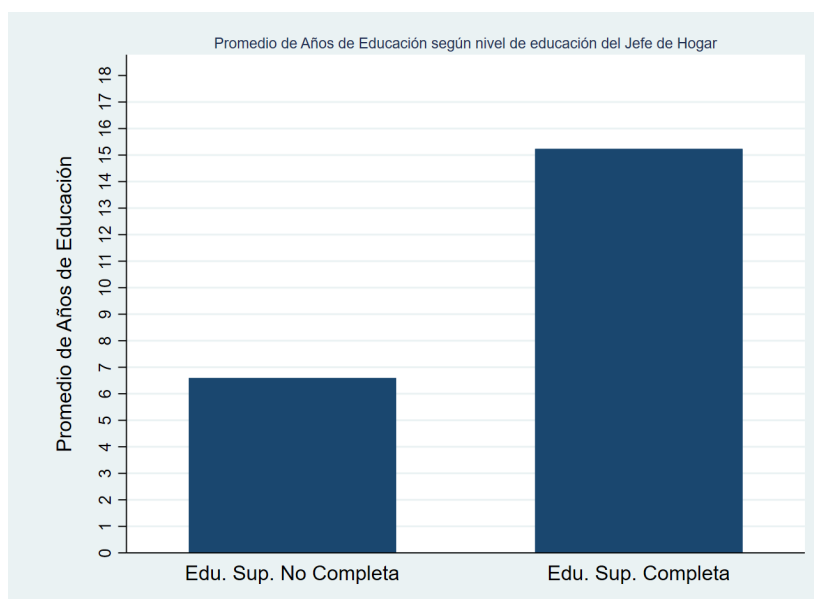
Graficando esta diferencia :

Usando gráfico de barras podemos visualizar mejor esta diferencia.

CODIGO:

```
label define jefedusup_labels 0 "Edu. Sup. No Completa" 1 "Edu. Sup. Completa"  
label values jefedusup jefedusup_labels  
  
* Crea el gráfico de barras con espacio adicional entre el eje y y el gráfico  
graph bar (mean) study_years, over(jefedusup) ///  
ylabel(0(1)18, labsize(*0.8) grid) ///  
title("Promedio de Años de Educación según nivel de educación del Jefe de Hogar", size(*0.5)) ///  
yttitle("Promedio de Años de Educación", size(*1.0) margin(medium)) ///  
legend(off)
```

RESULTADO:



EJERCICIO 2

Alizamos los logros educativos (años de educación promedion) en cada región del Perú

En este caso nos centraremos en los años 2004 y 2022. Es decir veremos como ha evolucionado la educación de las regiones desde el 2004 al 2022. Para eso es necesario trabajar con el módulo 300 de ENAHO, juntar las bases de datos para luego clasificarlos por cada departamento como se muestra en el siguiente código.

```
*primera pregunta:
*Variables geográficas (departamento, urbano/rural)
*Departamento (distinguir Lima Metropolitana de Lima Provincias)
destring ubigeo, generate(dpto)
replace dpto=dpto/10000
replace dpto=round(dpto)
label define dpto_num 1 "Amazonas"
label define dpto_num 2 "Ancash", add
label define dpto_num 3 "Apurimac", add
label define dpto_num 4 "Arequipa", add
label define dpto_num 5 "Ayacucho", add
label define dpto_num 6 "Cajamarca", add
label define dpto_num 7 "Callao", add
label define dpto_num 8 "Cusco", add
label define dpto_num 9 "Huancavelica", add
label define dpto_num 10 "Huanuco", add
label define dpto_num 11 "Ica", add
label define dpto_num 12 "Junin", add
label define dpto_num 13 "La_Libertad", add
label define dpto_num 14 "Lambayeque", add
label define dpto_num 15 "Lima_Metropolitana", add
label define dpto_num 16 "Lima_Provincias", add
label define dpto_num 17 "Loreto", add
label define dpto_num 18 "Madre_de_Dios", add
label define dpto_num 19 "Moquegua", add
label define dpto_num 20 "Pasco", add
label define dpto_num 21 "Piura", add
label define dpto_num 22 "Puno", add
label define dpto_num 23 "San_Martin", add
label define dpto_num 24 "Tacna", add
label define dpto_num 25 "Tumbes", add
label define dpto_num 26 "Ucayali", add
label values dpto dpto_num
```

Luego de ello podemos reutilizar el código para crear la variable de años de estudio del ejercicio anterior:

```
gen study_years = p301b
replace study_years= 0 if p301a == 1 | p301a == 2 // Sin nivel o Inicial
recode study_years (1=1) (2=2) (3=3) (4=4) if p301a==3 // Primaria incompleta
recode study_years (5=5) (6=6) if p301a==4 // Primaria completa
recode study_years (1=7) (2=8) (3=9) (4=10) if p301a==5 // Sec. incompleta
recode study_years (4=11) (5=11) (6=12) if p301a==6 // Sec. completa
recode study_years (1=12) (2=13) (3=14) (4=15) if p301a==7 // Sup. No Uni. incompleta
recode study_years (3=14) (4=15) (5=16) if p301a==8 // Sup. No Uni. completa
recode study_years (1=12) (2=13) (3=14) (4=15) (5=16) (6=17) if p301a==9 // Sup. Uni incompleta
recode study_years (4=15) (5=16) (6=17) (7=18) if p301a==10 // Sup. Uni completa
recode study_years (1=17) (2=18) if p301a==11 // Posgrado
```

Pero lo que nos importa analizar es el promedio de cada región y año. Para eso usamos sort y generamos nuevas variables.

```

sort dpto year
by dpto: egen educacion_promedio_2004 = mean(study_years) if year == "2004"
by dpto: egen educacion_promedio_2022 = mean(study_years) if year == "2022"

tab dpto, summarize(educacion_promedio_2022)
tab dpto, summarize(educacion_promedio_2004)
keep year study-years educación_promedio_2022 educación_promedio_2004 dpt

```

De esta forma obtenemos las promedios de cada una de las regiones por año.

PROMEDIO DE AÑOS DE ESTUDIOS POR DEPARTAMENTO EN EL AÑO 2004 :

. tab dpto, summarize(educacion_promedio_2004)			
ubicación geográfica	Summary of educacion_promedio_2004		
	Mean	Std. dev.	Freq.
Amazonas	3.9930093	0	3,006
Ancash	5.1699481	0	3,458
Apurimac	4.0784316	0	2,205
Arequipa	7.4551301	0	3,313
Ayacucho	4.3149066	0	3,221
Cajamarca	3.4818823	0	3,974
Callao	8.140976	0	1,302
Cusco	4.8423409	0	3,128
Huancavel	3.5885777	0	2,749
Huanuco	3.8656497	0	3,410
Ica	7.2355652	0	3,261
Junin	6.022305	0	3,509
La_Libert	5.8876858	0	3,237
Lambayequ	5.7397876	0	3,574
Lima_Metr	7.7114191	0	9,757
Lima_Prov	4.7384095	0	3,559
Loreto	5.7384334	0	2,184
Madre_de_	7.0152397	0	2,253
Moquegua	5.2706733	0	2,346
Pasco	5.2361407	0	4,277
Piura	5.2883983	0	3,052
Puno	4.4327044	0	3,180
San_Marti	7.3711977	0	2,602
Tacna	5.9987588	0	2,417
Tumbes	5.3300829	0	2,769
Total	5.6063189	1.3735387	81,743

PROMEDIO DE AÑOS DE ESTUDIOS POR DEPARTAMENTO EN EL AÑO 2022

:

```
. tab dpto, summarize(educacion_promedio_2022)
```

ubicaci@n geogr@fica	Summary of educacion_promedio_2022		
	Mean	Std. dev.	Freq.
Amazonas	5.0375032	0	4,133
Ancash	6.4776387	0	4,472
Apurimac	6.0401993	0	2,813
Arequipa	8.4842587	0	4,839
Ayacucho	5.7312217	0	3,316
Cajamarca	5.052382	0	4,429
Callao	8.6766834	0	3,469
Cusco	6.4639339	0	3,760
Huancavel	5.603189	0	3,261
Huanuco	5.4455194	0	3,928
Ica	8.5550413	0	4,751
Junin	7.2527719	0	4,512
La_Libert	6.7994595	0	5,182
Lambayequ	7.3991504	0	4,944
Lima_Metr	8.884676	0	14,707
Lima_Prov	5.681458	0	5,687
Loreto	7.0435243	0	1,887
Madre_de_	8.73316	0	2,692
Moquegua	6.5734768	0	2,792
Pasco	6.5453753	0	5,752
Piura	6.5702143	0	3,182
Puno	5.6390085	0	4,643
San_Marti	8.5765438	0	3,866
Tacna	7.4356332	0	2,914
Tumbes	6.2194219	0	4,326
Total	7.0114903	1.2983935	110,257

¿Qué regiones han avanzado más? Al parecer las regiones de Callao, Lima metropolitana y Arequipa han tenido un gran crecimiento en educación, mientras que otras como Amazonas, Madre de Dios y San Martín se han retrasado.

EJERCICIO 4

¿Qué región del Perú depende más del empleo independiente?

Para responder esta pregunta es necesario separar el empleo independiente del dependiente de la base de datos para esto creamos las variables `inglabde` (ingresos laborales dependientes) y `inglabinde` (ingresos laborales independientes). Sin embargo, esto no es todo debemos limpiar y corregir algunos valores en la base de datos, es decir limpiar los missing values. Es decir solo trabajaremos con la población activa, que está trabajando en un empleo. A la par obtenemos un ratio que nos permitirá hacer las comparaciones correspondientes. Posteriormente con la media de cada región de los trabajadores independientes y dependientes realizamos un loop, agregado a todo lo expresado antes.

```
foreach x of numlist 2013/2022{
    use "$dir_enaho/MODULO500/enaho01a-`x'-500", clear

    //Estimo los ingresos laborales dependientes e independientes

    "En primer lugar creamos las variables para ingresos laborales dependientes (inglabde) y para ingresos laborales independientes (inglabinde)."
    egen inglabde= rowtotal(i524a1 d529t i538a1 d540t i541a d544t)
    label var inglabde "Ingreso laboral dependiente"
    egen inglabinde= rowtotal(i530a d536 d543)
    label var inglabinde "Ingreso laboral independiente"

    "Elimino a todos los missing value en mi muestra de modo que no obtenga 0, es decir, no ocupados laboralmente."
    replace inglabde= . if i524a1== . & d529t== . & i538a1== . & d540t== . & i541a== . & d544t== .
    replace inglabinde= . if i530a== . & d536== . & d543== .

    //Obtener un ratio de ambos montos

    "genero una variable que nos de el ratio entre ambos"
    gen rating = inglabde / inglabinde //cuando me piden ratio me piden esto?
    label var rating "Ratio de ingresos"

    //Obtener el promedio para cada región

    "Primero debo establecer mis regiones. Para ello uso mi variable dominio y lo agrupo para recodificar la variable de modo que sea posible tener solo mis tres regiones naturales: costa, sierra y selva."
    recode dominio (1 2 3 8 = 1 "Costa") (4 5 6 = 2 "Sierra") (7 = 3 "Selva"), gen(regnat)
    label var regnat "Region Natural"

    * Promedio de ingresos laborales para trabajadores dependientes e independientes por región
    preserve
    collapse (mean) promde = inglabde prominde=inglabinde [iw=fac500a], by(regnat) //esta bien aca usar iw o no?
    gen year = `x'
    tempfile base`x'
    save `base`x''
    restore

    * Nacional
    preserve
    collapse (mean) promde = inglabde prominde=inglabinde [iw=fac500a]
    gen regnat = 4
    gen year = `x'
    tempfile base`x'`n'
    save `base`x'`n'
    restore
}
```

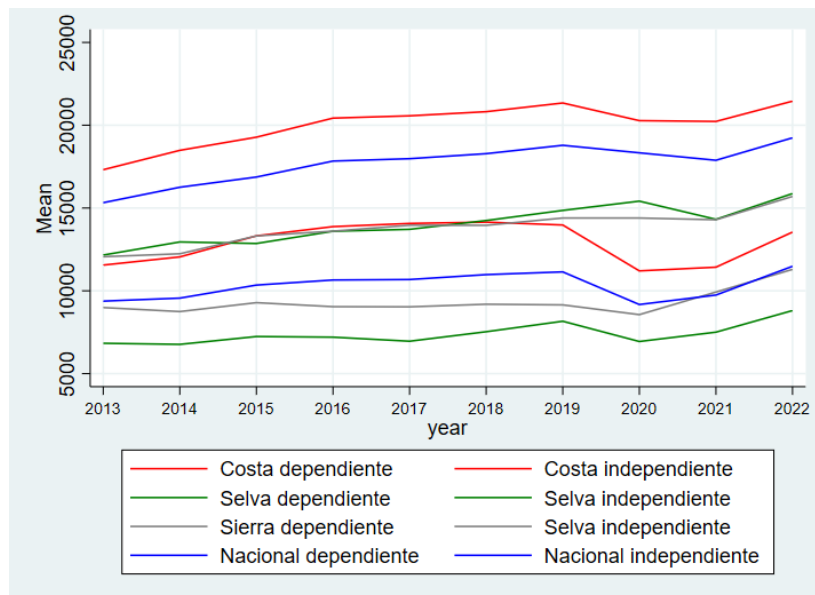
Adicionalmente para observar estas variables y disposiciones de mejor manera, usamos un diagrama. CODIGO:

```
clear
foreach x of numlist 2013/2022{
    append using `base`x'
    append using `base`x'`n'
}

* Con todo lo anterior ya tengo los promedios por región y también a nivel nacional.

//Generar un grafico con las características correspondientes a la tarea.
reshape wide promde prominde, i(year) j(regnat) // lo convierto a wide para poder hacer line
line promde1 promde2 promde3 promde4 promde5 promde6 promde7 promde8 year, xlabel(2013(1)2022, labsize(*0.8) grid) ///
xttitle("year") ytitle("Mean") ///
legend(order(1 "Costa dependiente" 2 "Costa independiente" 3 "Selva dependiente" 4 "Selva independiente" 5 "Sierra dependiente" 6 "Sierra independiente" 7 "Nacional dependiente" 8 "Nacional independiente")) ///
lcolor(red red green green gray gray blue blue)
```

RESULTADO:



EJERCICIO 5

Brecha salarial de género

La brecha salarial de género se ha convertido en un tema fundamental y de principal interés tanto en la investigación académica como en el desarrollo de políticas públicas. Esta desigualdad salarial tiene la finalidad, según los estudios de Gabriela Salas, de comparar el ingreso promedio entre hombres y mujeres asalariados (2019). Bajo esta misma línea, la distancia estimada entre los ingresos de hombre y mujeres pueden ser calculadas utilizando diferentes métodos y variables. Es así que de acuerdo con las variables utilizadas, son dos los indicadores predominantemente usados. El primero de ellos es la brecha salarial por género ajustada y la brecha salarial por género sin ajustar. Esta última toma en cuenta únicamente la variable sexo y población asalariada, mientras que la brecha salarial por género ajustada incluye el tipo de actividad prestada por hombres y mujeres, así como su jornada y tipo de contrato. Así, la brecha salarial sin ajustar da cuenta en términos amplios de la desigualdad por género en el ámbito laboral. Evidencia, pues, que la población femenina asalariada, percibe, en promedio, menores ingresos que la masculina asalariada. Por otro lado, el índice de la brecha salarial ajustada se podría comparar con el salario de hombres y mujeres que realizan un trabajo de igual valor o que trabajan en un mismo sector, o que cuentan con el mismo perfil académico. Este indicador, pues, busca medir la distancia entre ingresos de poblaciones con mayores semejanzas y hasta nos permite contar con data, aun frente a fenómenos como la segregación horizontal, la cual importa la segregación de mujeres a determinados sectores productivos “feminizados”. Es por ello que de acuerdo con las variables anteriormente expuestas, es posible deducir que no solo se toma en cuenta los promedios de ingreso laboral por sexo, ya que influyen factores que involucran un sesgo de elección, discriminación salarial, como asuntos culturales y sociales. Asimismo, la Organización Internacional del Trabajo ha investigado que la brecha salarial por género (sin ajustar) promedio en el mundo es del 23 % (OIT, 2016). En el Perú, la brecha salarial por género sin ajustar es del 29 %, lo cual significa que estamos 15.1 puntos porcentuales encima del promedio de brecha salarial de los países de la OCDE (OCDE, 2018). Por su parte, de acuerdo a estimaciones de la OIT, a falta de datos más actuales, al 2016, la brecha salarial ajustada en el Perú era del 17.5 % (OIT, 2018).

De un punto de vista más económico, la actual premio Nobel de economía, Claudia Goldin (2023) considera otros factores que amplían la brecha salarial de género. Uno de los más interesantes destacados por ella es el “Presentismo masculino”, que es la facilidad que tienen los hombres para estar casi todo el tiempo disponible para las empresas, mientras que las mujeres por diversos motivos biológicos y sociales no tienen tales facilidades.

En conclusión, la brecha salarial de género es un tema complejo y multidimensional que requiere un análisis riguroso para entender sus causas y efectos. Es por ello que para estimar la diferencia de ingresos laborales entre hombres y mujeres a nivel nacional, se necesitan datos que incluyan información detallada sobre los ingresos, la educación, la experiencia laboral y otros factores relevantes.

Estimando la diferencia de Ingresos laborales por género Checar la base de datos, para ver la diferencia.