www.desarrollosocial.cl

Ministerio de Desarrollo Social

Gobierno de Chile

Observatorio Social

Factor de expansión para el análisis de preguntas de orientación sexual e identidad de género

Serie documentos metodológicos Casen N° 34
18 mayo 2017

Casen 2015

Encuesta de Caracterización Socioeconómica Nacional

TABLA DE CONTENIDOS

1. Introducción

Atendiendo a las demandas planteadas desde la sociedad civil y, avanzando en la dirección de garantizar el ejercicio y reconocimiento pleno de derechos y libertades individuales, el Ministerio de Desarrollo Social ha incluido por primera vez preguntas en el cuestionario de la Encuesta Casen que buscan caracterizar, de forma anónima y confidencial, la identidad de género y la orientación sexual de las personas de 18 años o más de edad[1].

La pregunta sobre orientación sexual busca caracterizar a grupos de población considerando la atracción que expresan hacia personas de igual, distinto o ambos sexos.

La pregunta sobre identidad de género, en tanto, busca reconocer a las personas según éstas se identifiquen con el género masculino, femenino u otro, opción que puede concordar o no con su sexo biológico.

Ambas preguntas son auto-reportadas y no pueden ser respondidas por terceros. Se realiza a todas las personas de 18 años ó más, presentes en la vivienda al momento de realizarse la encuesta.

Cabe suponer que no todas las personas que forman parte del universo relevante (personas de 18 años o más) tienen la misma probabilidad de contestar estas preguntas, lo que podría introducir un sesgo en las estimaciones obtenidas a partir de los datos de la encuesta (más conocido como sesgo de selección). En términos simples, si las personas de 18 años o más que no responden las preguntas de diversidad sexual difieren de quienes sí responden, la distribución de respuestas y promedios estimados a partir de éstas darán cuenta únicamente de los atributos de quienes respondieron (no así, de aquellos que no lo hacen[2]).

Con el objeto de corregir eventuales sesgos, se construye un factor de expansión[3] específico para las preguntas de orientación sexual e identidad de género, que busca controlar por diferencias de características entre las personas que responden y aquéllas que no lo hacen, para que así las estimaciones obtenidas reflejen adecuadamente la distribución de atributos en todo el universo de población de interés (población de 18 años o más).

---

[1] Las preguntas de orientación sexual e identidad de género se encuentran presentes en el cuestionario Casen 2015 y son identificadas con las preguntas r21 y r22. El cuestionario y el Manual de Aplicación de la Encuesta Casen 2015 se encuentran publicados en:
http://observatorio.ministeriodesarrollosocial.gob.cl/casen-multidimensional/casen/casen_2015.php

[2] Las características diferentes deben están correlacionadas con las respuestas para la existencia de sesgo de selección.

[3] El factor de expansión corresponde a un valor de ponderación que permite expandir los resultados de la muestra al total de población correspondiente (considerando las proyecciones demográficas del Instituto Nacional de Estadísticas vigentes a la fecha de la Encuesta) y se interpreta como la cantidad de personas en la población que representa cada individuo en la muestra.

Este procedimiento es consistente con el aplicado para el análisis de preguntas incluidas en versiones anteriores de la Encuesta Casen, para las cuales sólo se registró respuesta de personas presentes en el momento de la Encuesta.

La utilización de este factor de expansión (identificado en la base de datos de la Encuesta Casen 2015 con la variable "*expr_div*") es requisito indispensable para realizar inferencias estadísticas sobre el total de población de 18 años o más con base en la información aportada por estas preguntas.

Es pertinente señalar, además, que este factor de expansión sólo es aplicable para generar estimaciones a nivel nacional y regional, no siendo factible su utilización para producir estimaciones en otros dominios de representación de la Encuesta.

2. Metodología aplicada

2.1. Diseño del factor de expansión

La construcción del factor de expansión se efectúa mediante un modelo *probit*[4]. De este modelo se obtienen nuevas probabilidades que corrigen el factor de expansión original a nivel hogar (factor de expansión regional, identificado en la base de datos con la denominación "expr"). Es decir, se genera una nueva variable en la base de datos que son los pesos (factor de expansión) asignados a cada observación (persona de 18 años o más) en la muestra y que aplican específicamente para el análisis de estas preguntas.

El objetivo de este factor de expansión es controlar el efecto que ejerce un conjunto de variables socioeconómicas y demográficas sobre la probabilidad de estar presente (y responder) al momento de aplicación de la encuesta. Bajo el supuesto que dicha probabilidad no se distribuye aleatoriamente, el factor de expansión asignará una ponderación condicionada a cada observación válida en la muestra (personas de 18 años o más presentes), de modo que la información recolectada pueda representar adecuadamente al total de población en el universo relevante (personas de 18 años o más)

Para este fin, el modelo utiliza un conjunto de variables independientes que influyen en la probabilidad de estar presente (y responder la encuesta) y cuya información procede de la misma encuesta.

4 Probit es una función de máxima verosimilitud que permite predecir probabilidades asegurando que la predicción se encontrará entre los valores [0,1]. Se asume que la distribución de los errores es normal.

2.2. Variables independientes incluidas en el modelo

La literatura especializada[5] entrega una clara idea del conjunto de variables que pueden ser consideradas en modelos de estas características. Sin embargo, la mayoría de la literatura se refiere a la no respuesta, que puede incluir: no contacto, razones ambientales, rechazo, y otros.

En este caso, se hace una opción por integrar variables socioeconómicas y demográficas que inciden en la presencia en el momento de la entrevista. Para tal propósito, se evaluó un amplio conjunto de variables potencialmente relevantes a incluir en el modelo, de las cuales –en función de su significancia estadística individual y mediante procedimiento *stepwise*- se mantuvieron las siguientes (Tabla 1):

Tabla 1: Variables independientes incluidas en el modelo.

Variable	Descripción
metrop	Variable dicotómica que toma el valor 1 si la persona reside en la región metropolitana y 0, si no.
sexo	Variable dicotómica que toma el valor 1 si el sexo de la persona es hombre y 0, si es mujer
asiste	Variable dicotómica que toma el valor 1 si la persona asiste a algún establecimiento educacional y 0, si no.
Jh	Variable dicotómica que toma el valor 1 si la persona es jefe de hogar y 0, si no.
Parejajh	Variable dicotómica que toma el valor 1 si el individuo es esposo/a o pareja del jefe de hogar y 0, si no.
Ind_tip	Variable dicotómica que toma el valor 1 si la vivienda en que reside la persona tiene un índice de tipo de vivienda aceptable y 0, si no (irrecuperable).
tedad183	Variable dicotómica que toma el valor 1 si la persona tiene 50 o más años de edad y 0, si no.
activ3	Variable dicotómica que toma el valor 1 si la condición de actividad de la persona es inactiva y 0, si no (no trabajó y no buscó trabajo activamente la semana anterior a la entrevista)
activ2	Variable dicotómica que toma el valor 1 si la persona se encuentra desocupada (no trabajó ni buscó trabajo activamente en semana anterior a la encuesta)
zona	Variable dicotómica que toma el valor 1 si la persona reside en zona rural y 0, si vive en zona urbana.
dific	Variable dicotómica que toma el valor 1 si la persona presenta alguna condición permanente o de larga duración

Fuente: Ministerio de Desarrollo Social.

[5] Véase, entre otros: Sarndal, C. (2001): "Estimation in the presence of nonresponse and frame imperfections".; Groves & Couper(1998): "Nonresponse in Household Interview Surveys"; EUSTAT (2007): "Estudio y Ajuste de la no Respuesta en las Encuestas de Hogares".

2.3. Ajuste por probabilidad de presencia al momento de la encuesta

La corrección del eventual sesgo de selección se realiza mediante un ajuste al factor de expansión de aquellas personas que responden la pregunta. Este ajuste corresponde al inverso de la probabilidad de estar presente durante la encuesta y responderla (esta consideración debe hacerse al momento de realizar la predicción del modelo).

La probabilidad de estar presente en la encuesta se calcula mediante un modelo probit, en el cual la variable dependiente es una variable dicotómica que toma valores 1 (persona está presente al momento de la entrevista y responde pregunta sobre orientación sexual) y 0 (persona no responde la entrevista).

El modelo probit determina la probabilidad de responder (rdiversidad=1) y un conjunto de variables que podrían influir en ella, tomando como referencia la respuesta a la pregunta r21 (orientación sexual). Nótese que basta con que la persona responda a la pregunta r21, independientemente de la orientación sexual reportada, para que la variable dependiente tome el valor 1. De lo contrario, se infiere que la persona no respondió.

Los resultados obtenidos se sintetizan en Anexo 1.

2.4. Síntesis del procedimiento aplicado

A continuación se presenta un resumen de los pasos ejecutados[6]:

i. Modelo probit

- **a.** Se realiza un modelo probit con las variables del modelo. La variable dependiente es "responde la pregunta de orientación sexual".
- b. La estimación del modelo se realizó considerando que el factor de expansión pondera a las variables como pesos de muestreo[7]. Es necesario calcular la predicción del modelo (rdiv). Esta predicción va a tomar valores que están en el intervalo [0,1].
- **c.** Generar una variable dicotómica (presp) que tome el valor 1 si la predicción toma valores 0.5 o superior. De lo contrario, la variable dicotómica tomará el valor 0.

ii. Generar pesos corregidos y ajustados a la población

- **a.** El modelo considera dos tipos de pesos. El primero corresponde a dividir el factor de expansión regional por la predicción de la regresión del probit (rdiv). Cabe destacar que se consideran sólo a las personas que respondieron la pregunta r21 (rdiversidad=1).
- **b.** El segundo peso divide el factor de expansión regional por veintiles de la predicción de la regresión del probit (rsatisf). Para construir los veintiles es necesario tomar en cuenta sólo a las personas que respondieron la pregunta r21 (rdiversidad=1).

[6] La sintaxis de programación para software Stata se expone en Anexo 2.

Anexo 1: Regresión y efecto marginal del modelo probit.

Tabla 2: Resultados de la regresión y efecto marginal del modelo probit.

VARIABLES	(1) Probit Responde pregunta de orientación sexual (r21)	(2) Efecto marginal Responde pregunta de orientación sexual (r21)
Región Metropolitana (metrop)	-0 ,0233**	-0 ,00925**
	(1 ,02265e-02)	(4 ,05783e-03)
Sexo	-0 ,701***	-0 ,273***
	(1 ,04008e-02)	(3 ,90338e-03)
Asiste	-0 ,105***	-0 ,0416***
	(2 ,00279e-02)	(7 ,98333e-03)
Jefe de hogar (jh)	1 ,261***	0 ,462***
	(1 ,28060e-02)	(4 ,02775e-03)
Pareja del jefe (parejajh)	0 ,748***	0 ,279***
	(1 ,41130e-02)	(4 ,73883e-03)
Indice de tipo de vivienda (ind_tip)	-0 ,410***	-0 ,155***
	(5 ,65183e-02)	(1 ,97164e-02)
50 años o más (tedad183)	-0 ,0871***	-0 ,0346***
	(1 ,06022e-02)	(4 ,20562e-03)
Inactivo (activ3)	0 ,261***	0 ,103***
	(1 ,08726e-02)	(4 ,24387e-03)
Desocupado (activ2)	0 ,321***	0 ,123***
	(2 ,24055e-02)	(8 ,22273e-03)
Zona	0 ,0355***	0 ,0140***
	(1 ,01876e-02)	(4 ,02599e-03)
Dific	-0 ,0327**	-0 ,0130**
	(1 ,44765e-02)	(5 ,75347e-03)
Constant	0 ,0806	
	(5 ,76987e-02)	
Observations	199,920	199,920
Pseudo R-squared	0 ,157	0 ,157

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Fuente: Ministerio de Desarrollo Social, Encuesta Casen 2015.

Anexo 2: Sintaxis de programación aplicada.

```
use "xxx.dta",  clear /*usar base de datos casen 2015*/

*Dummy traslado
gen traslado=.
replace traslado=0 if o25a_hr==0 & o25a_min==0
replace traslado=1 if ((o25a_hr>0 & o25a_hr<=3) | (o25a_min>0 & o25a_min<59))

*Dummy jornada
gen jornada_tarde=.
replace jornada_tarde=0 if e9==1 | e9==3 | e9==4
replace jornada_tarde=1 if e9==2 | e9==5 | e9==6

*Comunas grandes
bysort comuna: egen auxcom=sum(expr)
gen comgrande=(auxcom>50000)

*Dummy zona
recode zona 1=0
recode zona 2=1
label define zona 0 "Urbana" 1 "Rural", replace

*DUMMY DE SEXO (MUJER = 0)
recode sexo 2=0
label define sexo 1 "Hombre" 0 "Mujer", replace

*Variable de asistencia escolar
recode asiste 2=0

*Jefe de hogar o pareja
gen jh=.
replace jh=1 if (pco1==1)
replace jh=0 if pco1>1 & pco1<=15
label var jh "Jefe de hogar"
label define jh 0 "Otro miembro del hogar" 1 "Jefe de hogar o pareja"
label values jh jh

*Pareja del jefe
gen parejajh=.
replace parejajh=1 if pco1==2
replace parejajh=0 if (pco1==1 | (pco1>2 & pco1<=14))
label var parejajh "Pareja del jefe"
label define parejajh 0 "Otro miembro del hogar" 1 "Pareja del jefe"
label values parejajh parejajh

*Tramos de edad
gen tedad18=.
replace tedad18 = 1 if edad>=18 & edad<30
replace tedad18 = 2 if edad>=30 & edad<50
replace tedad18 = 3 if edad>=50 & edad<.
```

```
label var tedad18 "Tramos de edad"
label define tedad18 1 "18 a 29 años" 2 "30 a 49 años" 3 "50 años o más"
label value tedad18 tedad18
tab tedad18, gen(_tedad18)

*Actividad
tab activ, gen(_activ )

*Desocupados
gen desoc=(activ==2)
replace desoc=. if activ==.

*Condiciones permanentes
gen dific=.
replace dific=1 if s31c1<=6
replace dific=0 if s31c1==7
label var dific "Condiciones permanentes"
label define dific 1 "Presenta dificultad" 0 "No presenta Dificultades"
label values dific dific

*Región metropolitana
gen metrop=(region==13)

*Nacionalidad
gen extranjeros=(r1a==3)
replace extranjeros=. if r1a==9

*Indice del tipo de vivienda
gen ind_tip=.
replace ind_tip=1 if (v1<=6 | v1==8)
replace ind_tip=0 if (v1==7 | v1==9)
label var ind_tip "índice de tipo de vivienda"
label define ind_tip 1 "aceptable" 0 "irrecuperable"
label value ind_tip ind_tip

*Norte
gen norte=(region==1 | region==2 | region==3 | region==4 | region==15)

*Ingreso
gen ypcautcorh=yautcorh/numper
xtile veiautnac = ypcautcor [w=expr], nq(20)


*RESPONDE PREGUNTA DE DIVERSIDAD (15 AÑOS O MÁS)*
gen rdiversidad=.
replace rdiversidad=0 if (r21==.) & edad>=15
replace rdiversidad=1 if (r21>=1 & r21<=9) & edad >=15
label var rdiversidad "Responde pregunta de diversidad"
label define rdiversidad 0 "No responde" 1 "Sí responde"
label values rdiversidad rdiversidad
```

```
gen pesosdiv= expr/rdiv if rdiversidad==1

*Probando modelo
local var zona sexo asiste jh parejajh _tedad181 _tedad182 _tedad183 _activ1 _activ2 _activ3
metrop ind_tip dific extranjeros veiautnac
sum rdiversidad `var'
corr rdiversidad `var'
xi:stepwise, pr(.05):  probit rdiversidad zona sexo asiste jh parejajh i.tedad18 i.activ dific
metrop extranjeros ind_tip [pw=expr] , r
estimates store m1
xi:dprobit rdiversidad metrop sexo asiste jh parejajh ind_tip _tedad183 _activ3 _activ2 zona
dific  [pw=expr], r
estimates store m2
xi:stepwise, pr(.05):  probit rdiversidad zona sexo asiste jh parejajh i.tedad18 i.activ dific
metrop extranjeros ind_tip  [pw=expr] , r
estimates store m3
outreg2 [m1 m2 m3] using "C:\probitr21_final.xls", label replace bfmt(f) sdec(7) sfmt(e)
decmark( ,) addstat(Pseudo R-squared, `e(r2_p)') fmt(g)  /*donde guardes el excel*/
predict rdiv

probit rdiversidad metrop sexo asiste jh parejajh ind_tip _tedad183 _activ3 _activ2 zona dific
[pw=expr] , r
estimates store e1
dprobit rdiversidad metrop sexo asiste jh parejajh ind_tip _tedad183 _activ3 _activ2 zona dific
[pw=expr], r
estimates store e2
outreg2 [e1 e2] using "C:\probitr21_final_2.xls", label replace bfmt(f) sdec(7) sfmt(e) decmark(
,) addstat(Pseudo R-squared, `e(r2_p)') fmt(g)  /*donde guardes el excel*/


*** Con predicción media por veintil ***

xtile v20_rdiv = rdiv [w=pesosdiv] if rdiversidad==1 & edad>=15, nq(20)
egen meanv20_rdiv = mean (rdiv) if rdiversidad==1 & edad>=15, by(v20_rdiv)
gen pv20_rdiv = expr/meanv20_rdiv if edad>=15 & rdiversidad==1

** Ajuste a población regional a partir de predicción media por veintil ***
egen pregorig= sum(expr) if edad>=15, by(region)
egen pregcorr= sum(pv20_rdiv) if pv20_rdiv>0 & pv20_rdiv!=., by(region)
gen factreg= pregorig/pregcorr if pv20_rdiv>0 & pv20_rdiv!=.
gen expr_r2div2=pv20_rdiv*factreg
label var expr_r2div2 "Pesos satisfacción con la vida"

gen presp = (rdiv>=0.5)
label var presp "Probabilidad estimada de haber respondido pregunta de diversidad>=0,5"
label define presp 0 "No" 1 "Sí"
label values presp presp

preserve
keep expr_r2div2 folio o
```

```
sort folio o
save "xxx.dta", replace
restore
```