dosr: Acelerando y Estandarizando el Análisis de Encuestas Complejas en el Sector Público

Palabras clave: srvyr, encuestas complejas, sector público, automatización, reproducibilidad.

Gabriel Sotomayor López

Abstract

Las instituciones del sector público dependen del análisis de encuestas con diseños muestrales complejos para informar políticas públicas. Este proceso suele ser manual, con posibilidad de errores y difícil de estandarizar. Para abordar estos desafíos, dentro de la División de Observatorio Social (DOS) del Ministerio de Desarrollo Social y Familia de Chile se desarrolló dosr, un paquete de R que encapsula el flujo de trabajo completo. Sus funciones de alto nivel (`obs_prop`, `obs_media`, etc.), integradas con el ecosistema `srvyr`, automatizan el cálculo de estimaciones para múltiples encuestas simultáneamente, siendo aplicable a cualquier conjunto de encuestas complejas que el usuario prepare como objetos `tbl_svy`. Sus características principales son la implementación programática de criterios de fiabilidad estadística institucionales y el cálculo automático de pruebas de significancia. Finalmente, dosr genera reportes en Excel listos para difusión. Esta charla presenta dosr como un caso de estudio sobre cómo R puede aumentar la eficiencia y estandarizar la calidad de la evidencia estadística en el servicio público.

Introducción: El Desafío del Análisis de Encuestas en el Sector Público

El análisis riguroso de encuestas nacionales es una tarea fundamental para el monitoreo social en el sector público. Este proceso requiere no solo el cálculo de estimadores, sino también la evaluación de su calidad estadística para asegurar la robustez de la evidencia. Tradicionalmente, este flujo de trabajo no está estandarizado, abriendo espacio a inconsistencias y requiriendo la aplicación manual de criterios de fiabilidad (basados en grados de libertad y errores estándar o coeficiente de variación) y la programación ad-hoc de pruebas de significancia.

Para enfrentar estos problemas, se desarrolló una solución interna utilizando R, lo cual se suma a una solución previa desarrollada en STATA, con el objetivo de automatizar, estandarizar y acelerar el proceso completo, desde el cálculo hasta el reporte final.

La Solución: El Paquete dosr

La solución implementada fue el paquete de R dosr. Esta herramienta proporciona una interfaz de alto nivel que abstrae la complejidad del análisis de encuestas, permitiendo a los analistas enfocarse en la interpretación de resultados. Integrado con srvyr, el paquete ofrece cinco funciones principales, obs_prop(), obs_media(), obs_cuantil(), obs_total y obs_ratio() que con una sola llamada ejecutan el ciclo completo de análisis.

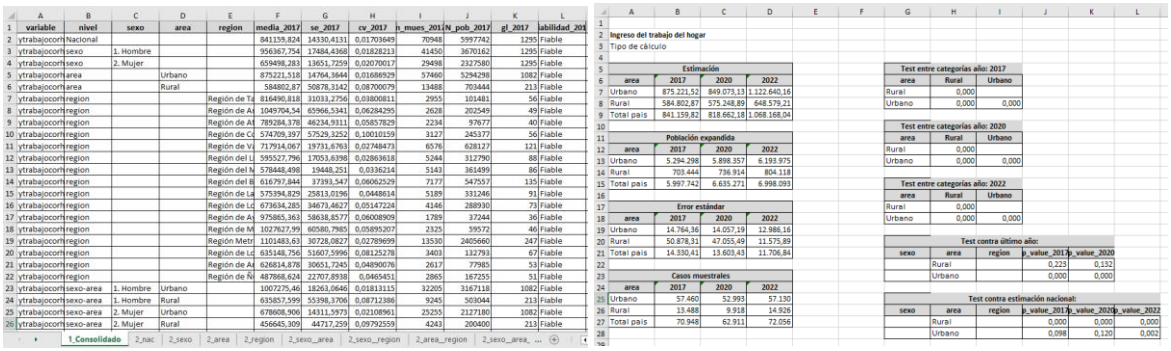
```
# Ejemplo: Media de ingresos por sexo y área para tres años, con tests de significancia

obs_media(designs = list("2017" = casen_2017d, "2020" = casen_2020d, "2022" = casen_2022d),
        var = "ytrabajocorh",
        des = c("sexo", "area","region"),
        parallel =T,
        formato = T,
        multi_des = T,
        filt = "pco1==1 & nucleo!= 0",
        sig = TRUE)
```

Características y Flujo de Trabajo

El diseño de dosr encapsula las mejores prácticas del análisis de encuestas en una herramienta simple y potente:

- Suite Completa de Estimadores: Calcula proporciones, medias, cuantiles, totales y ratios a través de una API coherente.
- Manejo de Múltiples Encuestas: Facilita el análisis comparativo y de tendencias al aceptar una lista de objetos tbl_svy. Para garantizar comparaciones válidas, es responsabilidad del usuario asegurar la estandarización de variables y etiquetas entre los diferentes diseños.
- Desagregaciones Flexibles: El usuario puede solicitar cruces por múltiples variables (multi_des = TRUE) o solo desagregaciones simples (multi_des = FALSE).
- Implementación de Criterios de Calidad: El motor interno aplica automáticamente el flujo de decisión para la fiabilidad de las estimaciones, añadiendo una columna fiabilidad ("Fiable", "Poco Fiable", "No Fiable") a todos los resultados.
- Pruebas de Significancia Automatizadas: Con sig = TRUE, el paquete calcula p-values para tres tipos de comparaciones: entre categorías, contra el total nacional y a través del tiempo.
- Generación de Reportes: La salida principal es un archivo Excel con una hoja de datos consolidados y múltiples hojas con tablas de presentación formateadas, listas para su difusión (ver Figura 1).



variable	nivel	sexo	area	region	media_2017	se_2017	cv_2017	n_mues_2017	N_pob_2017	gl_2017	abilidad_201
ytrabajocorh	Nacional				841159,824	14330,4131	0,01703649	70948	3997742	1295	Fiable
ytrabajocorh	sexo	1. Hombre			556367,754	17484,4368	0,01828213	41450	3670162	1295	Fiable
ytrabajocorh	sexo	2. Mujer			659498,283	13651,7259	0,02070017	29498	2327580	1295	Fiable
ytrabajocorh	area		Urbano		875221,518	14764,3644	0,01686929	57460	5294298	1082	Fiable
ytrabajocorh	area		Rural		584802,87	50878,3142	0,08700079	13488	703444	213	Fiable
ytrabajocorh	region			Región de Ta	816490,818	31033,2756	0,03800811	2955	101481	56	Fiable
ytrabajocorh	region			Región de Ar	1049704,54	65966,5341	0,06284295	2628	202549	49	Fiable
ytrabajocorh	region			Región de At	789284,378	46234,9311	0,05857829	2234	97677	40	Fiable
ytrabajocorh	region			Región de Co	574709,397	57529,3252	0,10010159	3127	245377	56	Fiable
ytrabajocorh	region			Región de Vi	717914,067	19731,6763	0,02748473	6576	628127	121	Fiable
ytrabajocorh	region			Región del L	595527,796	17053,6398	0,02863618	5244	312790	88	Fiable
ytrabajocorh	region			Región del N	578448,498	19448,251	0,0336214	5143	361499	86	Fiable
ytrabajocorh	region			Región del B	616797,844	37393,547	0,06062529	7177	547557	135	Fiable
ytrabajocorh	region			Región de La	575394,829	25813,0196	0,0448614	5189	331246	91	Fiable
ytrabajocorh	region			Región de Lo	673634,285	34673,4627	0,05147224	4146	288930	73	Fiable
ytrabajocorh	region			Región de Ar	975865,363	58638,8577	0,06008909	1789	37244	36	Fiable
ytrabajocorh	region			Región de M	1027627,99	60580,7985	0,05895207	2325	59572	46	Fiable
ytrabajocorh	region			Región Metr	1101483,63	30728,0827	0,02789699	13530	2405660	247	Fiable
ytrabajocorh	region			Región de Lo	635148,756	51607,5996	0,08125278	3403	132793	67	Fiable
ytrabajocorh	region			Región de Ar	626814,878	30651,7245	0,04890076	2617	77985	53	Fiable
ytrabajocorh	region			Región de Ñ	487868,624	22707,8938	0,0465451	2865	167255	51	Fiable
ytrabajocorh	sexo-area	1. Hombre	Urbano		1007275,46	18263,0646	0,01813115	32205	3167118	1082	Fiable
ytrabajocorh	sexo-area	1. Hombre	Rural		635857,599	55398,3706	0,08712386	9245	503044	213	Fiable
ytrabajocorh	sexo-area	2. Mujer	Urbano		678608,906	14311,5973	0,02108961	25255	2127180	1082	Fiable
ytrabajocorh	sexo-area	2. Mujer	Rural		456645,309	44717,259	0,09792559	4243	200400	213	Fiable

1_Consolidado | 2_nac | 2_sexo | 2_area | 2_region | 2_sexo_area | 2_sexo_region | 2_area_region | 2_sexo_area_ ...

Ingreso del trabajo del hogar

Tipo de cálculo

Estimación

area	2017	2020	2022
Urbano	875.221,52	849.073,13	1.122.640,16
Rural	584.802,87	575.248,89	648.579,21
Total país	841.159,82	818.662,18	1.068.168,04

Población expandida

area	2017	2020	2022
Urbano	5.294.298	5.898.357	6.193.975
Rural	703.444	736.914	804.118
Total país	5.997.742	6.635.271	6.998.093

Error estándar

area	2017	2020	2022
Urbano	14.764,36	14.057,19	12.986,16
Rural	50.878,31	47.055,49	11.575,89
Total país	14.330,41	13.603,43	11.706,84

Casos muestrales

area	2017	2020	2022
Urbano	57.460	52.993	57.130
Rural	13.488	9.918	14.926
Total país	70.948	62.911	72.056

Test entre categorías año: 2017

area	Rural	Urbano
Rural	0,000	
Urbano	0,000	0,000

Test entre categorías año: 2020

area	Rural	Urbano
Rural	0,000	
Urbano	0,000	0,000

Test entre categorías año: 2022

area	Rural	Urbano
Rural	0,000	
Urbano	0,000	0,000

Test contra último año:

sexo	area	region	p_value_2017	p_value_2020
	Rural		0,225	0,132
	Urbano		0,000	0,000

Test contra estimación nacional:

sexo	area	region	p_value_2017	p_value_2020	p_value_2022
	Rural		0,000	0,000	0,000
	Urbano		0,098	0,120	0,002

Figura 1: Ejemplo de hoja de reporte generada por dosr, mostrando estimaciones y pruebas de significancia.

Impacto y Desafíos Futuros

La adopción de dosr, junto con otras herramientas desarrolladas por el equipo, ha tenido un impacto directo en la eficiencia y calidad del trabajo de análisis. Las tareas se realizan de manera más eficiente y estandarizada, entregando una mejor herramienta para asegurar que todas las estimaciones publicadas son sometidas a los mismos estándares de fiabilidad. Esto no solo mejora la reproducibilidad, sino que también facilita la incorporación de nuevos analistas al equipo.

El paquete dosr es un ejemplo de cómo el desarrollo de herramientas de código abierto dentro del sector público puede mejorar sustancialmente los flujos de trabajo. Los desafíos y planes futuros se centran en:

- Implementar una suite de pruebas unitarias (testthat) para garantizar la estabilidad a largo plazo.
- Generar un sitio web del paquete (pkgdown) para mejorar la documentación y facilitar su adopción.
- Abstraer los criterios de fiabilidad para que puedan ser personalizados por el usuario, aumentando la utilidad del paquete en otros contextos institucionales.