Proyecto Final: Ejercicio 1

Alejandra Lelo de Larrea 124433, Diego A. Estrada 165352, Victor Quintero 175897

Se estima el total de votos para las elecciones de julio 2012 en México por candidato a partir de muestras de distintos tamaños de secciones nominales. Para ello, se utiliza Muestreo Aleatorio Simple (SI) y Muestreo con Probabilidades Proporcionales (PP) al listado nominal con el fin de comparar los estimadores para ambos métodos y los distintos tamaños de muestra. Como se están estimando totales, se decidió utilizar el estimador de Narain-Horvitz-Thompson (NTH) en ambos diseños de muestreo. Además, se utilizó el método de máxima entropía para la selección de muestras en PP. En cuanto a la estimación de las varianzas, para muestreo SI se utiliza la estimación de NHT y para muestreo PP el estimador de Sen-Yates-Grundy (SGY). Para efectos de comparabilidad, y para no correr el riesgo de obtener valores negativos, al calcular el DEFF en PP se reestimó la varianza del muestreo SI con el estimador SGY. Cabe mencionar que se tienen 27 secciones nominales para las cuales no se tienen datos debido a que se están trabajando con datos históricos y en las elecciones del 2012 dichas secciones nominales no existían; por ello, se decidió eliminar estas observaciones de la muestra.

1. Comparación de métodos y tamaños de muestra

El porcentaje de secciones utilizado, así como el listado nominal y el porcentaje de listado nominal utilizado en cada muestra para cada uno de los métodos se encuentra en la tabla 1. Un aspecto a resaltar es que el porcentaje de secciones nominales utilizados en la muestra es menor a 1 % en 4 de los 5 casos. El efecto de utilizar probabilidades proporcionales al listado nominal se puede notar en el hecho de que el porcentaje del listado nominal considerado en cada muestra bajo muestreo aleatorio simple, es menor que el de muestreo con probabilidades proporcionales para todos los casos. De hecho, la mayoría de las muestras utilizan menos del 1 % del listado nominal para realizar las estimaciones.

Tabla 1: Listado nominal por tamaño de muestra y tipo de diseño muestral.

No. Secciones	%	Aleatorio Simple		Prob. Proporcionales	
en muestra	secciones	Total Votantes	% List. Nom	Total Votantes	% List. Nom
50	0.08	66,564	0.0838	81,808	0.1030
100	0.15	139,902	0.1762	216,382	0.2725
250	0.38	279,805	0.3523	501,219	0.6312
500	0.75	603,591	0.7601	1,055,722	1.3294
6500	9.78	7,637,875	9.6180	13,821,946	17.4052

La figura 1 compara el total de votos estimado para cada candidato bajo los dos métodos y para los distintos tamaños de muestra. Por su parte, la tabla 2 resume los resultados de las estimaciones. Un vistazo general tanto de las gráficas como de las métricas, sugiere que el muestreo con probabilidades proporcionales da mejores estimaciones que el muestreo aleatorio simple en la mayoría de los casos; dado que en este ejercicio en particular se puede conocer el total de votos verdadero por cada candidato, es fácil verificar esta afirmación. A pesar de utilizar un porcentaje tan pequeño, los resultados obtenidos con muestreo con probabilidades proporcionales son satisfactorios para la mayoría de las muestras. En contraste, los resultados obtenidos con muestreo aleatorio simple para tamaños de muestra de 50 y 100 sobreestiman el total de votos para la mayoría de los candidatos, pero son cercanos al verdadero valor en muestras grandes.

De la figura 1 y de las columnas 3 a 5 de la tabla 2 se observa, como era de esperarse, que la diferencia entre el total estimado con SI y el verdadero total de votos decrece (en valor absoluto) conforme se incrementa el tamaño de muestra para los candidatos AMLO, EPN y GQT; sin embargo, para JVM, Nulos y No Registrdos esto no se cumple. Para el muestreo PP de manera general se observa que para el tamaño de muestra 250 la diferencia (en valor absoluto) entre el total estimado y el verdadero valor se incrementa drásticamente para algunos candidatos (AMLO, JVM y Nulos). Además, en el SI la mejor estimación se obtiene para el mayor tamaño de muestra; en contraste, bajo PP en el 50 % de los casos (AMLO, JVM, Nulos) se obtiene la estimación más cercana para la muestra de tamaño 500 y en el otro 50 % de los casos (EPN, GQT y No Registrados) se obtiene

la mejor estimación con el mayor tamaño de muestra. Con esto se puede concluir que utilizar muestreo con probabilidades proporcionales puede ayudar realizar un muestreo más eficiente y menos costos si la variable de interés esta fuermente correlacionada con la variable utilizada para obtener las probabilidades de inclusión.

De manera general, en las columnas 6 a 9 de la tabla 2 se puede notar que los errores estándar y, por ende, los intervalos de confianza, son más pequeños bajo PP que bajo SI. En general, la amplitud del intervalo disminuye conforme aumenta el tamaño de muestra. Además, los verdaderos votos totales quedaron contenidos dentro de los intervalos de confianza a excepción de 4 casos; de éstos tres corresponden a SI (GQT con tamaño de muestra 250 y votos nulos con tamaños de muestra 250 y 6500) y únicamente uno a PP (AMLO con tamaño de muestra 6500).

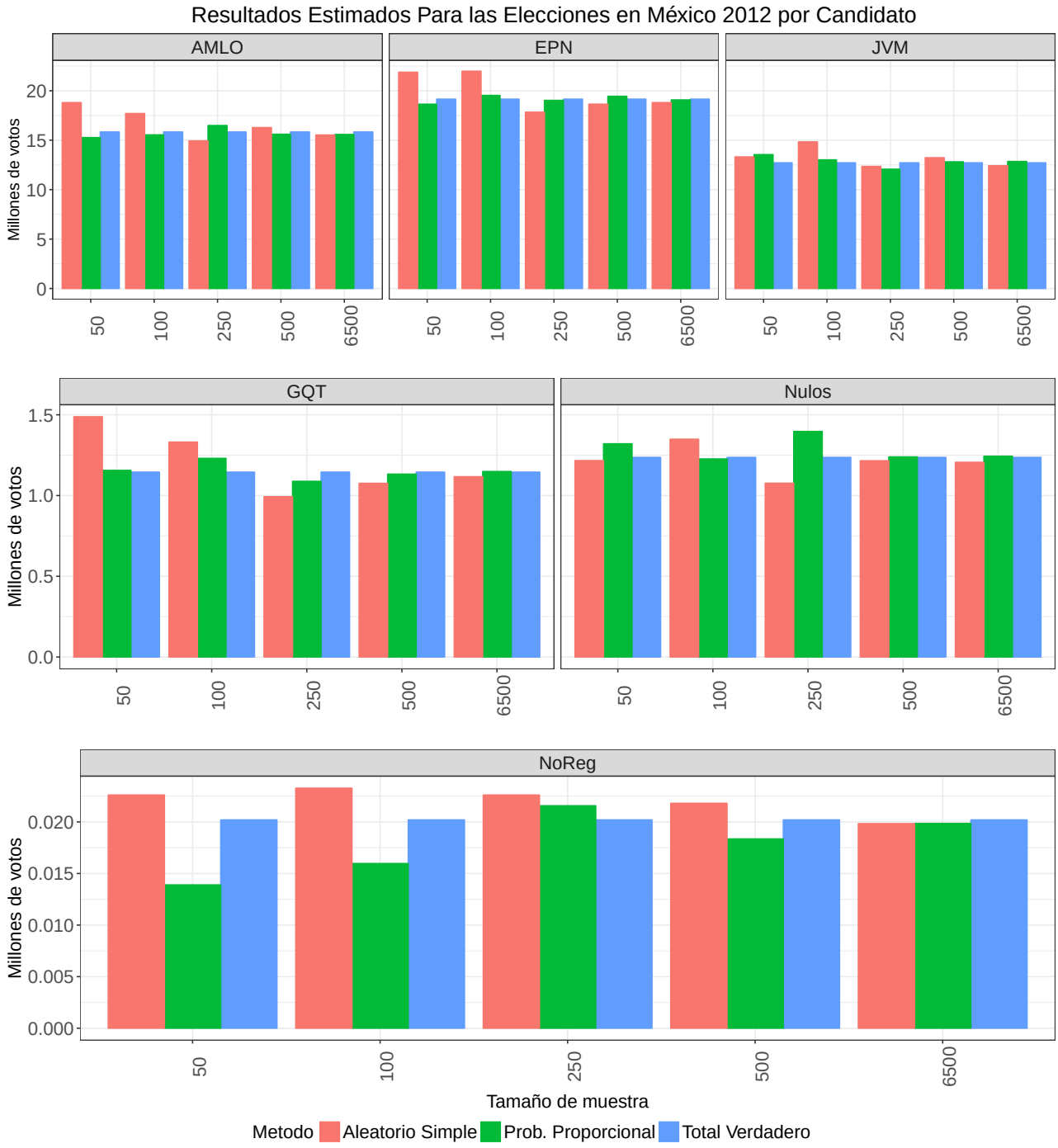

Figura 1: Total de votos por candidato para los distintos tamaños de muestra y los distintos métodos de estimación.

Tabla 2: Métricas para el total de votos estimado por candidato, por tamaños de muestra y por método de estimación.

a) AMLO

Método	n	Total	Tot. Estim.	Diferencia	Error Est.	Límite Inf.	Límite Sup.	Amplitud	CVE	DEFF
SI	50	15,832,258	18,800,430	-2,968,172	2,411,862	14,073,267	23,527,593	9,454,326	0.13	1.00
	100	15,832,258	17,700,702	-1,868,444	2,197,614	13,393,458	22,007,946	8,614,488	0.12	1.00
	250	15,832,258	14,928,110	904,148	916,385	13,132,028	16,724,193	3,592,165	0.06	1.00
	500	15,832,258	16,276,640	-444,382	774,121	14,759,390	17,793,890	3,034,500	0.05	1.00
	6500	15,832,258	15,524,946	307,312	188,520	15,155,454	15,894,439	738,985	0.01	1.00
PP	50	15,832,258	15,268,567	563,691	1,299,437	12,721,718	17,815,416	5,093,698	0.09	0.29
	100	15,832,258	15,538,873	293,385	827,188	13,917,615	17,160,131	3,242,516	0.05	0.14
	250	15,832,258	16,483,162	-650,904	545,125	15,414,737	17,551,587	2,136,850	0.03	0.35
	500	15,832,258	15,601,015	231,243	390,392	14,835,861	16,366,169	1,530,308	0.03	0.25
	6500	15,832,258	15,576,396	255,862	95,641	15,388,943	15,763,849	374,906	0.01	0.26

b) EPN

Método	n	Total	Tot. Estim.	Diferencia	Error Est.	Límite Inf.	Límite Sup.	Amplitud	CVE	DEFF
SI	50	19,151,414	21,873,551	-2,722,137	2,644,200	16,691,014	27,056,089	10,365,075	0.12	1.00
	100	19,151,414	21,980,599	-2,829,185	2,290,804	17,490,705	26,470,492	8,979,787	0.10	1.00
	250	19,151,414	17,834,211	1,317,203	859,169	16,150,271	19,518,152	3,367,881	0.05	1.00
	500	19,151,414	18,645,510	505,904	680,038	17,312,660	19,978,361	2,665,701	0.04	1.00
	6500	19,151,414	18,804,378	347,036	184,329	18,443,100	19,165,657	722,557	0.01	1.00
PP	50	19,151,414	18,641,863	509,551	781,036	17,111,061	20,172,665	3,061,604	0.04	0.09
	100	19,151,414	19,526,965	-375,551	683,494	18,187,340	20,866,589	2,679,249	0.04	0.09
	250	19,151,414	19,026,630	124,784	391,154	18,259,983	19,793,278	1,533,295	0.02	0.21
	500	19,151,414	19,442,065	-290,651	311,160	18,832,203	20,051,927	1,219,725	0.02	0.21
	6500	19,151,414	19,073,325	78,089	72,383	18,931,456	19,215,194	283,738	0.00	0.15

c) JVM

Método	n	Total	Tot. Estim.	Diferencia	Error Est.	Límite Inf.	Límite Sup.	Amplitud	CVE	DEFF
SI	50	12,714,460	13,313,757	-599,297	2,224,312	8,954,186	17,673,329	8,719,142	0.17	1.00
	100	12,714,460	14,827,712	-2,113,252	2,248,673	10,420,395	19,235,029	8,814,634	0.15	1.00
	250	12,714,460	12,347,539	366,921	777,197	10,824,261	13,870,818	3,046,556	0.06	1.00
	500	12,714,460	13,227,322	-512,862	925,509	11,413,357	15,041,286	3,627,928	0.07	1.00
	6500	12,714,460	12,419,296	295,164	155,834	12,113,868	12,724,724	610,856	0.01	1.00
PP	50	12,714,460	13,541,511	-827,051	1,008,768	11,564,363	15,518,660	3,954,297	0.07	0.21
	100	12,714,460	13,012,384	-297,924	879,953	11,287,707	14,737,061	3,449,353	0.07	0.15
	250	12,714,460	12,068,119	646,341	445,487	11,194,981	12,941,257	1,746,276	0.04	0.33
	500	12,714,460	12,811,434	-96,974	294,989	12,233,266	13,389,601	1,156,335	0.02	0.10
	6500	12,714,460	12,846,507	-132,047	80,761	12,688,217	13,004,796	316,579	0.01	0.27

d) GQT

Método	n	Total	Tot. Estim.	Diferencia	Error Est.	Límite Inf.	Límite Sup.	Amplitud	CVE	DEFF
SI	50	1,145,187	1,488,024	-342,837	268,326	962,115	2,013,933	1,051,819	0.18	1.00
	100	1,145,187	1,331,110	-185,923	200,499	938,139	1,724,081	785,942	0.15	1.00
	250	1,145,187	992,814	152,373	67,485	860,545	1,125,083	264,538	0.07	1.00
	500	1,145,187	1,075,792	69,395	59,380	959,410	1,192,174	232,763	0.06	1.00
	6500	1,145,187	1,116,841	28,346	18,270	1,081,033	1,152,650	71,617	0.02	1.00
PP	50	1,145,187	1,156,062	-10,875	84,592	990,264	1,321,860	331,596	0.07	0.10
	100	1,145,187	1,230,413	-85,226	89,182	1,055,620	1,405,207	349,587	0.07	0.20
	250	1,145,187	1,088,177	57,010	54,882	980,609	1,195,744	215,135	0.05	0.66
	500	1,145,187	1,132,673	12,514	36,661	1,060,819	1,204,527	143,708	0.03	0.38
	6500	1,145,187	1,148,955	-3,768	9,287	1,130,753	1,167,157	36,404	0.01	0.26

e) Nulos

Método	n	Total	Tot. Estim.	Diferencia	Error Est.	Límite Inf.	Límite Sup.	Amplitud	CVE	DEFF
SI	50	1,236,474	1,216,749	19,725	145,656	931,269	1,502,228	570,959	0.12	1.00
	100	1,236,474	1,349,062	-112,588	126,825	1,100,489	1,597,635	497,146	0.09	1.00
	250	1,236,474	1,076,590	159,884	54,050	970,653	1,182,527	211,874	0.05	1.00
	500	1,236,474	1,215,552	20,922	52,865	1,111,938	1,319,166	207,229	0.04	1.00
	6500	1,236,474	1,206,172	30,302	13,524	1,179,666	1,232,678	53,012	0.01	1.00
PP	50	1,236,474	1,320,653	-84,179	135,363	1,055,346	1,585,959	530,613	0.10	0.86
	100	1,236,474	1,226,819	9,655	63,038	1,103,267	1,350,371	247,105	0.05	0.25
	250	1,236,474	1,396,902	-160,428	100,554	1,199,821	1,593,984	394,163	0.07	3.46
	500	1,236,474	1,239,383	-2,909	34,175	1,172,401	1,306,366	133,965	0.03	0.42
	6500	1,236,474	1,243,530	-7,056	8,539	1,226,794	1,260,265	33,471	0.01	0.40

f) No Registrados

Método	n	Total	Tot. Estim.	Diferencia	Error Est.	Límite Inf.	Límite Sup.	Amplitud	CVE	DEFF
SI	50	20,197	22,606	-2,409	7,006	8,875	36,337	27,462	0.31	1.00
	100	20,197	23,271	-3,074	6,291	10,942	35,600	24,659	0.27	1.00
	250	20,197	22,606	-2,409	3,964	14,837	30,376	15,539	0.18	1.00
	500	20,197	21,808	-1,611	2,156	17,584	26,033	8,449	0.10	1.00
	6500	20,197	19,834	363	639	18,582	21,086	2,504	0.03	1.00
PP	50	20,197	13,899	6,298	4,459	5,159	22,639	17,480	0.32	0.41
	100	20,197	15,969	4,228	4,031	8,068	23,869	15,801	0.25	0.41
	250	20,197	21,568	-1,371	4,381	12,980	30,155	17,174	0.20	1.22
	500	20,197	18,352	1,845	2,053	14,328	22,375	8,047	0.11	0.91
	6500	20,197	19,853	344	615	18,647	21,058	2,411	0.03	0.93

Por su parte, en las columnas 10 y 11 de dicha tabla se encontró que al utilizar un diseño de muestreo con PP se obtiene un CVE menor (o en su defecto igual) al obtenido bajo SI. Cabe destacar que las únicas excepciones en las que el CVE bajo SI fue menor que bajo propabilidades proporcionales, corresponden a los casos de Nulos (con $n = 250$) y a No Registrados (con $n = 50, 250, 500$); posiblemente esto tenga que ver con que el tamaño del listado nominal no tiene una relación tan estrecha con el total de votos anulados ni con el total de votos por candidatos no registrados. Además, para ambos métodos el menor CVE se obtiene con el mayor tamaño de muestra; ésto no puede generalizare para el valor del DEFF en el muestreo con probabilidades proporcionales. Estos resultados se deben principalmente a que, con probabilidades proporcionales, sí se toma en consideración la cantidad de electores que podrían votar en cada una de las secciones nominales al asignarles diferentes factores de expansión; mientras que en el muestreo SI se considera el mismo factor de expansión para todas las secciones nominales en la muestra.

2. Resultados Electorales

Es difícil confiar por completo en todas las estimaciones para las muestras de menor tamaño. Por ejemplo, para el caso de AMLO con PP y tamaño de muestra 50, el total estimado es de aproximadamente 15 millones, mientras que la amplitud del intervalo de confianza es de una tercera parte de dicha estimación. Un segundo ejemplo se tiene para la estimación del total de votos para candidatos no registrados bajo SI con una muestra de tamaño 100, donde la amplitud del intervalo de confianza es mayor al total estimado. Sin embargo, a partir del tamaño de muestra 250 la proporcion entre la estimación del total y la amplitud del intervalo de confianza disminuye considerablemente conforme aumenta el tamaño de muestra; esto da señales de que es posible confiar más en las estimaciones para las muestras grandes.

De esta manera, utilizando el menor CVE como criterio de selección, para dar una estimación de los resultados electorales se utilizarían las estimaciones del tamaño de muestra 6500 con probabilidades proporcionales para todos los candidatos. Los totales estimados se pueden consultar en la tabla 2 de la sección anterior. Este criterio, es inconsistente únicamente para las estimaciones de JVM y para el total de votos nulos ya que bajo PP, pero con tamaño de muestra 500, se obtienen estimaciones más cercanas al verdadero valor. Sin embargo, en la vida real no es factible tener distintas muestras para poder hacer la comparación ni el verdadero valor de la variable de interés para verificar la calidad de las estimaciones.

La tabla 3 muestra el ranking de los candidatos bajo las estimaciones de las distintas muestras. Se puede notar que independientemente del método y del tamaño de muestra, el primer lugar es para EPN, el segundo lugar para AMLO y el tecer lugar para JVM. Junto con la información de la tabla 2, se puede concluir que Gabriel Quadri es el candidato más sensible a cambios en el tamaño de muestra puesto que alterna entre el cuarto y quinto lugar del ranking con los votos nulos (aunque en su mayoría GQT queda en quinto lugar).

Tabla 3: Ranking de candidatos por tamaños de muestra y por método de estimación.

Lugar	Aleatorio Simple					Prob. Proporcionales				
	n=50	n=100	n=250	n=500	n=6500	n=50	n=100	n=250	n=500	n=6500
1	EPN	EPN	EPN	EPN	EPN	EPN	EPN	EPN	EPN	EPN
2	AMLO	AMLO	AMLO	AMLO	AMLO	AMLO	AMLO	AMLO	AMLO	AMLO
3	JVM	JVM	JVM	JVM	JVM	JVM	JVM	JVM	JVM	JVM
4	GQT	Nulos	Nulos	Nulos	Nulos	Nulos	GQT	Nulos	Nulos	Nulos
5	Nulos	GQT	GQT	GQT	GQT	GQT	Nulos	GQT	GQT	GQT
6	No Reg.	No Reg.	No Reg.	No Reg.	No Reg.	No Reg.	No Reg.	No Reg.	No Reg.	No Reg.

Se puede concluir que, en general, todos los diseños de muestreo especificados en este ejercicio funcionaron para el fin último de estimar al ganador de las elecciones presidenciales de México en 2012. Sin embargo, es importante destacar que el muestreo con probabilidades proporcionales resultó mucho mejor que el muestreo SI en términos de precisión y exactitud, sobre todo para tamaños de muestra pequeños; ésto se traduce en estimadores más estables y cercanos al verdadero valor.