¿Cómo y cuánto fallan los sondeos electorales?

Pedro Delicado*, Dept. d'Estadística i I.O., UPC
Frederic Udina, Dept. d'Economia i Empresa, UPF

Resumen:
En este trabajo se presenta una metodología sencilla de evaluación de las predicciones de los sondeos electorales. Tanto la descripción gráfica como las medidas numéricas propuestas se basan en métodos de simulación. Se presta especial atención al problema de la estimación (sesgada) de la distribución de escaños entre partidos políticos mediante la ley d'Hondt y a la estimación de diferencias. Se estudia el origen del sesgo en la estimación y se sugieren métodos para su reducción. En ambos casos se analiza el problema de la elección previa del tamaño muestral para garantizar un margen de error dado. Los resultados y las predicciones de las elecciones catalanas de octubre de 1999 y las elecciones generales de marzo de 2000 ilustran el trabajo.

Palabras clave: Elección del tamaño muestral, estadística pública, ley d'Hondt, métodos multivariantes, Monte Carlo.

How and how much do the pre-election polls fail?

Abstract:
In this paper we present a simple methodology to evaluate pre-election polls forecasting. The graphic and numerical proposed tools are based on simulation procedures. We pay special attention to the bias in the estimation of the seats allocation by d'Hont rule, and to the estimation of proportion differences. The origin of the estimation bias is studied and some bias reduction techniques are proposed. We analyze the previous choice of a sample size warranting a given margin of error in seats allocation and difference estimation. The work is illustrated with two elections examples: Catalan regional parliament elections (October, 1999) and Spanish parliament elections (March, 2000).

Keywords: Sample size choice, public statistics, d'Hondt rule, multivariate methods, Monte Carlo.

*Dirección de contacto: Pedro Delicado, Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Edifici U, C/ Pau Gargallo, 5; 08028 Barcelona.

1 Introducción

A raíz de los malos pronósticos de los sondeos publicados ante las elecciones al Parlament de Catalunya de octubre de 1999 (en adelante Parlament'99) y las elecciones generales de marzo de 2000 (en adelante Congreso'00) quisimos analizar desde el punto de vista probabilístico el problema de la predicción de resultados en el contexto de la Ley Electoral española, que incorpora como mecanismo de reparto de escaños la ley d'Hondt.

Cuando se realiza un sondeo electoral se obtiene una muestra aleatoria de la población que conforma el censo. Sabido es que los principales problemas para el análisis de esta muestra residen en la dificultad de obtener de ella respuestas fiables, incluso en el supuesto que los entrevistados sepan realmente lo que votarán en el momento decisivo. Todas las encuestas publicadas utilizan algún mecanismo de imputación de datos faltantes para paliar el problema de la falta de respuesta. En su publicación, pocas de ellas aportan datos sobre el mecanismo utilizado, por lo que no entraremos a discutir este aspecto. Nos limitaremos a discutir problemas imputables únicamente al muestreo. Aún en el supuesto de que todos los entrevistados respondan fiablemente, quedan interesantes problemas por analizar.

Aunque hay trabajos interesantes sobre metodología de encuestas electorales, citemos Bernardo (1984), no tenemos conocimiento de ningún trabajo que estudie el problema estadístico de la estimación de escaños asignados mediante una regla como la ley de Hondt.

Consideraremos un modelo teórico de los sondeos electorales. En él, K partidos se disputan un total de N escaños repartidos en C circunscripciones, con N_i, $(i = 1, \ldots, C)$ escaños en cada una de ellas. La muestra será una muestra aleatoria estratificada de un total de n elementos, repartidos entre las circunscripciones a razón de n_i elementos en la circunscripción c_i.

En este artículo nos situamos en este marco para tratar diversos problemas relacionados con los sondeos electorales. En la sección 2 se aborda el problema de visualizar simultáneamente los resultados derivados de diferentes sondeos. La sección 3 muestra las dificultades prácticas provocadas por las peculiaridades matemáticas de la ley d'Hondt. La falta de coherencia entre los datos estadísticos publicados en los medios de comunicación y las conclusiones que se manifiestan en esos mismos medios es el tema de la sección 4, que considera los casos de la predicción de diferencias entre proporciones y de las horquillas de escaños. En la sección 5 se dan recomendaciones sobre cómo elegir el tamaño muestral en un sondeo electoral para conseguir objetivos definidos en términos del margen de error permitido en la estimación de diferencias de proporciones o de asignación de escaños. Hemos dejado para los apéndices el tratamiento completo y razonado de las propuestas apuntadas en las secciones anteriores. Así, el apéndice A analiza la regla d'Hondt desde una perspectiva matemática, el apéndice B discute los problemas probabilísticos relacionados con la estimación de múltiples proporciones y de sus diferencias, así como el problema de la elección del tamaño muestral. Finalmente, recogemos en el apéndice C los datos disponibles sobre los sondeos electorales utilizados en el texto.

2 Visualización de sondeos electorales

En los días siguientes a las elecciones aparece la discusión de por qué los sondeos preelectorales se equivocaron. Esto fue especialmente notorio en las elecciones tanto al Parlament'99 como al Congreso'00. Hemos desarrollado una metodología para analizar y visualizar el error cometido por estos sondeos. No entramos en la discusión del porqué de los errores, sino en mostrar cómo y cuánto se equivocaron.

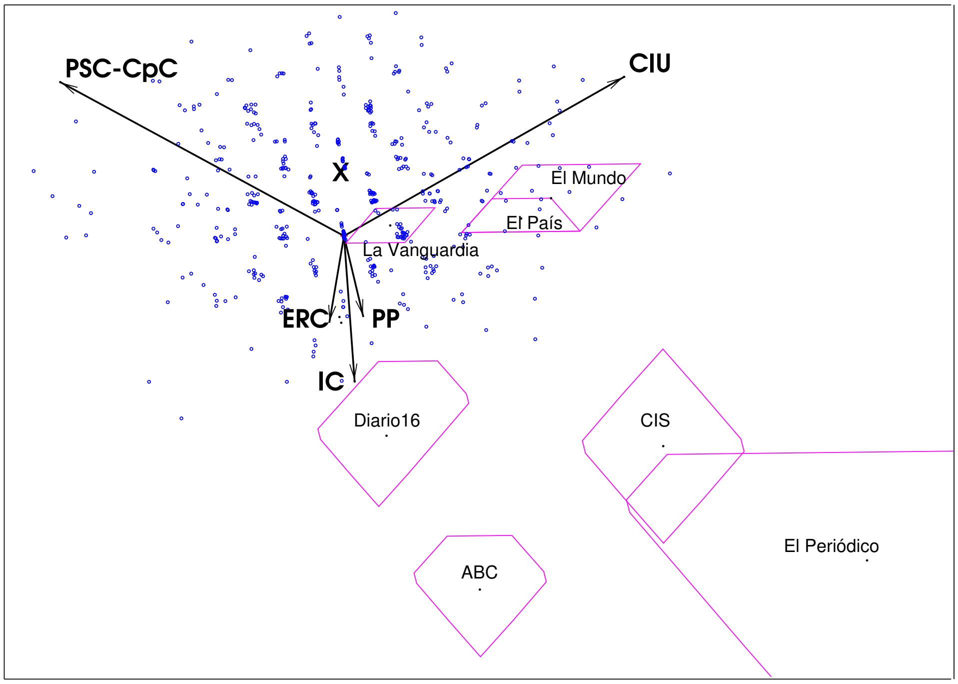

Figura 1: Gráfico basado en componentes principales en el que se representan los sondeos preelectorales publicados las semanas anteriores a las elecciones al Parlament de Catalunya 1999 por diversos medios de comunicación. Cada polígono representa las horquillas de escaños pronosticadas por el sondeo. Los puntos representan los parlamentos predichos por cada uno de 2000 sondeos teóricos obtenidos por simulación. Las flechas representan las direcciones que favorecen a cada uno de los partidos, con origen situados en el parlamento real. Se marca con una X el parlamento promedio de los sondeos simulados.
Fuente: Elaboración propia.

La metodología, cuyo resultado gráfico se muestra en la figura 1, consiste en simular gran número de sondeos teóricos tomando como parámetros los más comunes entre los sondeos publicados. Por ejemplo, para las elecciones al Parlament'99, tomando como proporciones poblacionales las que dieron las urnas en cada provincia catalana, el tamaño muestral se fijó en 800 para Barcelona y 400 para cada una de las provincias restantes. Con estos datos, simulamos en el ordenador $B = 2000$ sondeos utilizando las distribuciones multinomiales apropiadas. Sobre los resultados de cada sondeo se aplica la ley d'Hondt para calcular los escaños de cada partido. Esto nos da una nube de puntos (2000 en este caso, de los que sólo dibujamos 500 para mayor legibilidad del gráfico) en un espacio de 6 dimensiones (5 partidos y 'otros'). El análisis de componentes principales permite representar lo más fielmente posible esta nube en un gráfico plano. En el mismo gráfico representamos las direcciones correspondientes a cada partido, tomando como origen el parlamento real proyectado sobre el plano del gráfico. También proyectamos sobre el mismo gráfico las predicciones de escaños de distintos sondeos publicados en los medios en fechas próximas a los comicios. Para ello, calculamos todos los parlamentos posibles dentro de la horquilla dada, proyectamos los puntos correspondientes sobre el plano de las componentes principales y dibujamos la envolvente convexa de estos puntos para no complicar la lectura del gráfico. En los casos en que el sondeo publicado se basa en un tamaño muestral distinto, corregimos la posición y el tamaño del polígono correspondiente mediante un factor $\sqrt{n_b/n}$, donde n_b és el tamaño utilizado en los sondeos teóricos.

El primer hecho que destaca del gráfico obtenido es la importancia del sesgo: la distancia entre el parlamento real, calculado a partir de las proporciones realmente salidas de las urnas, y el parlamento promedio obtenido por los 2000 sondeos simulados (marcado con una X en el gráfico). Dedicaremos la sección 3 a analizar el origen de este sesgo pero subrayemos aquí que la presencia de este sesgo no puede ser ignorada al establecer predicciones de escaños mediante sondeos como los que se realizan en la práctica.

También es destacable la gran diferencia entre los tamaños aparentes y entre las posiciones de los distintos sondeos publicados que se incluyen en el gráfico (los datos técnicos de dichos sondeos se listan en la sección C). Queda claro que los errores de las predicciones no pueden atribuirse al azar muestral en la mayoría de los casos, y sorprende especialmente que las desviaciones se dan en dirección contraria a la que debería producirse dado el sesgo que produce el muestreo.

Hemos aplicado la misma metodología a los sondeos publicados ante las elecciones generales al Parlamento español de Marzo del 2000. El resultado puede verse en la figura 2. El cálculo se basa en 2000 sondeos simulados, de los cuales sólo se visualizan en la nube de puntos 500, para mayor claridad del gráfico. El tamaño muestral utilizado es de $N = 15.000$ con asignación en parte fija y en parte proporcional, tal como se especificaba en el único sondeo publicado con una ficha técnica lo suficientemente precisa (en El País).

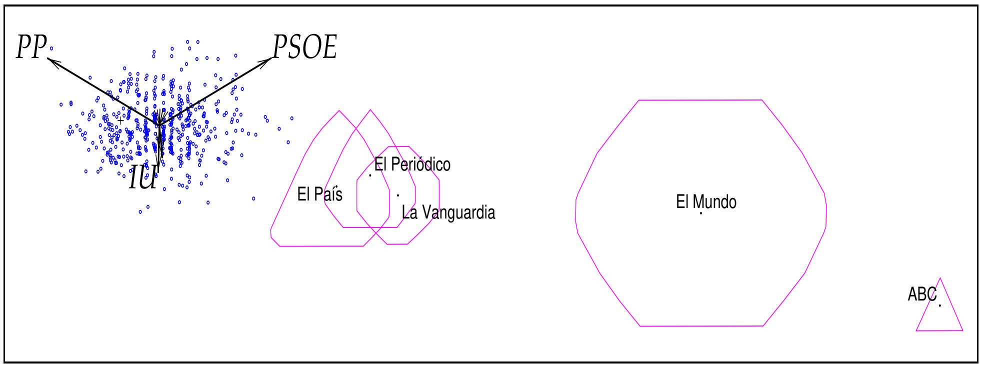

Figura 2: Gráfico basado en componentes principales en el que se representan los sondeos preelectorales publicados las semanas anteriores a las elecciones al Parlamento español 2000 por diversos medios de comunicación. La lectura del gráfico es la misma de la figura anterior.
Fuente: Elaboración propia.

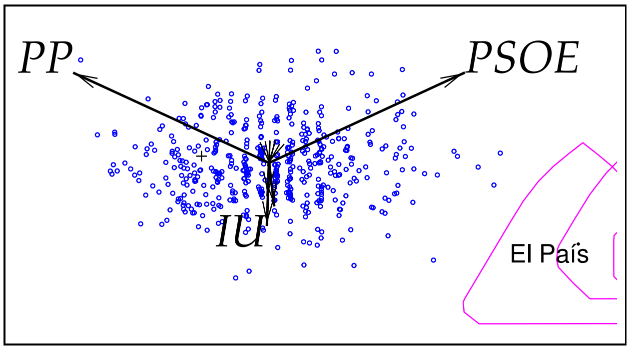

Figura 3: Ampliación de la nube de puntos de la figura 2. A la izquierda del origen común de las flechas se puede distinguir el parlamento real marcado con una cruz, que se distingue mejor en la ampliación de la parte inferior de la figura.
Fuente: Elaboración propia.

Las dos primeras componentes principales utilizadas para la construcción del gráfico acumulan un 81% de la varianza. Sólo se han rotulado las flechas de los tres partidos principales. Las de todos los partidos menores coinciden en dirección prácticamente con la de IU. También aquí aparece el sesgo en la estimación de la asignación de escaños: puede distinguirse a la izquierda del origen una pequeña cruz que indica la posición del parlamento real, a partir de cuyas proporciones se han simulado los sondeos (véase el detalle en la figura 3. Destacamos que el tamaño del sesgo es comparable al radio de las horquillas con que predicen el parlamento la mayoría de sondeos publicados, por lo que no es despreciable en absoluto. Los polígonos convexos que representan a los sondeos publicados se han calculado en la misma forma que para la figura anterior. Destaca claramente la infravaloración del voto del PP y, de forma peculiar, la estrechez de las horquillas dadas por el diario ABC, que de hecho cubrían únicamente tres composiciones del parlamento posibles.

3 Los problemas de la regla d'Hondt

La regla d'Hondt es la fórmula adoptada por la legislación electoral española para el reparto de escaños. Para repartir N escaños entre K partidos que han obtenido votos respectivos $(f_1, f_2, \ldots; f_K)$ se forman los *cocientes de Hondt* $(f_i, f_i/2, f_i/3, \ldots, f_i/N$ para cada partido) y se atribuye un escaño a cada uno de los N mayores cocientes.

En el apéndice A analizamos al detalle el funcionamiento de la regla, así como su análisis matemático. Nos interesa resaltar aquí que, en el contexto de un sondeo electoral que pretende predecir la distribución de escaños, la regla d'Hondt es una función aleatoria que depende de las proporciones muestrales. El hecho destacable es que la estimación de los escaños es sesgada: el promedio de las predicciones realizadas a través de muchos sondeos, no coincidiría con el resultado final. Dicho de otra forma, la predicción de escaños de un sondeo debería ser corregida para ser creíble.

Para entender este efecto, analizamos algunos de los casos más simples. En una circunscripción electoral como la de Ceuta hay un sólo escaño en juego. El partido que tenga más votos se lo adjudicará. Si sólo hubiera dos partidos en liza, lo que en realidad es prácticamente cierto, y si el primer partido obtuviera el 50% o más de los votos, se quedaría con el escaño. Si en tal circunscripción la proporción de votantes del PP fuera próxima al 50%, pongamos $p = 0.55$, la variabilidad muestral nos podría llevar a predecir que el escaño es para el PP con bastante facilidad.[1]

[1] Si $p = 0.45$ y el tamaño muestral es $n_i = 116$, la probabilidad de asignar el escaño incorrectamente seria del 14%. Este tamaño muestral fue el utilizado en Ceuta por el sondeo publicado por El País, si bien la proporción de votos del PP fue del 71%, con una probabilidad de asignación errónea prácticamente nula.

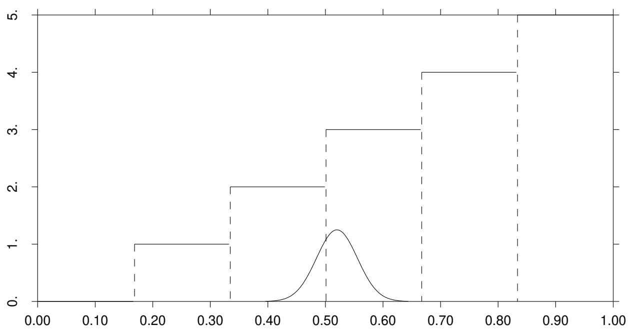

Figura 4: Distribución de escaños en una circunscripción en la que dos partidos se disputan cinco escaños. En el eje horizontal el porcentaje de votos de uno de los partidos, en el vertical el número de escaños que se le otorgan. En el supuesto de que el 52% de votantes sean para este partido, los sondeos de tamaño $n_i = 199$ obtendrían proporciones muestrales distribuidas según la campana que aparece en la parte inferior del gráfico.
Fuente: Elaboración propia.

En Cáceres los cinco escaños se los disputan prácticamente a solas el PP y el PSOE. La figura 4 muestra el número de escaños que corresponden a uno de los partidos en función del porcentaje de votos que obtenga. La campana de la parte inferior del gráfico visualiza la distribución de las proporciones muestrales que se obtendrían en sondeos de tamaño $n_i = 199$ si la proporción real de votos fuera del 52%[2] En tal situación, el 28% de los sondeos darían una predicción errónea de dos escaños, el resto acertarían los tres escaños para el partido más votado.

Más interesante es una situación en que tres partidos se disputan cierto número de escaños. La visualizamos en la figura 5 con los datos correspondientes a La Rioja en las elecciones al Congreso'00 (PP, PSOE y IU debían repartirse 4 escaños). A partir del análisis que se detalla en el apéndice A construimos un triángulo en el que se pueden representar todas las combinaciones posibles de resultados en porcentajes y en escaños. Cada punto del triángulo representa una combinación de tres porcentajes que suman 100. Así el punto R marcado en el gráfico representa el resultado que dieron las urnas, 58.2%, 37.5% y 4.3% para PP, PSOE e IU respectivamente, tras excluir los votos obtenidos por los partidos que no superaron el umbral del 3%. En el triángulo se han dibujado también los polígonos que corresponden a todas las combinaciones de votos que dan una misma asignación de escaños: la combinación R otorga 3 escaños al PP, 1 al PSOE y ninguno a IU. Si realizamos un sondeo de tamaño $n_j = 167$[3] las proporciones que

[2]Este fue el porcentaje obtenido por el PP en las elecciones al Congreso'00. El tamaño muestral fue el utilizado por el sondeo de El País.

[3]Este fue el tamaño muestral utilizado por el sondeo publicado en el País.

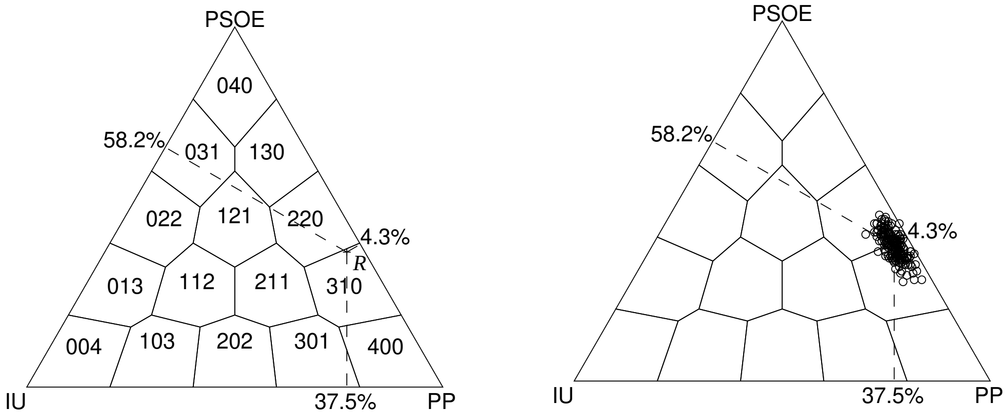

Figura 5: Triángulos donde se representan las proporciones de votos de tres partidos en una circunscripción con cuatro escaños en juego. Los porcentajes se representan en coordenadas triangulares: el punto R corresponde al 58.2% para el PP, 37.5% para el PSOE y 4.3% para IU (resultados en La Rioja, Congreso'00; porcentajes de votos, tras excluir los de los partidos que no superaron el umbral del 3%). Los polígonos corresponden a las combinaciones de porcentajes que dan igual reparto de escaños, los rótulos de cada polígono dan el número de escaños para cada partido, en el orden anterior. El gráfico de la derecha incorpora las predicciones de 200 sondeos de tamaño $n = 167$ simulados por ordenador.
Fuente: Elaboración propia.

obtendremos serán similares a las de la población pero nunca iguales. En la misma figura, a la derecha, hemos dibujado un punto para cada una de las proporciones obtenidas en 200 sondeos simulados por ordenador. Puede observarse que sólo la mitad de los sondeos caen en la zona correcta (en la que está el punto R, el verdadero resultado), con lo que en la mitad de los sondeos la asignación de los escaños sería incorrecta.

Este sesgo se produce siempre y es el exponente de la dificultad en la predicción del número de escaños que obtendrán los partidos. Cuando este efecto se reproduce en cada una de las circunscripciones electorales, hasta 52 en el caso de las elecciones al congreso, el sesgo de las estimaciones resulta ser importante, como se pone de manifiesto en las figuras 1 y 3.

Un procedimiento que ayuda a paliar en parte el sesgo de los sondeos consiste en usar técnicas de Monte Carlo del modo siguiente. Una vez llevado a cabo el sondeo, se usan los datos obtenidos como si fueran los verdaderos y se replica este primer sondeo mediante simulación tantas veces como se desee. Se calcula el parlamento medio de los obtenidos en las simulaciones. La diferencia entre ese promedio y el parlamento que se derivaba del primer sondeo (el único real) es una estimación del sesgo que conlleva el proceso de estimación. Si al parlamento estimado originalmente le restamos ese sesgo, tendremos una estimación *corregida* de sesgo.

4 Horquillas y otras confianzas

La publicación de los sondeos electorales o de otro tipo en los periódicos se acompaña de una ficha técnica en la que se puede leer algo como "El error muestral para un nivel de confianza del 95.5% es del ±2.8% en el supuesto más desfavorable (p=q=0.5)". Ello debe interpretarse del siguiente modo: si el muestreo se repitiese 1000 veces y cada vez se construyera un intervalo de confianza para la proporción de interés p, aproximadamente en 955 ocasiones (o más) dicho intervalo contendría el verdadero (y desconocido) valor de p, incluso si ese valor es igual a 0.5, que el más difícil de estimar. Pero, como discutimos a continuación, a veces se dan interpretaciones incorrectas de las fichas técnicas.

En la segunda parte de esta sección discutimos el nivel de confianza de las horquillas de escaños que se dan en los sondeos electorales, que no son más que intervalos de confianza (de confianza no especificada, eso sí). Los aspectos más técnicos los hemos recogido en el apéndice B.

4.1 Hablemos de precisión con precisión

Una semana antes de las elecciones a la Generalitat de Catalunya, El País titulaba una página con la frase *"Pujol se despega de Maragall [...] al que supera en 3.5 puntos [porcentuales]"*, basándose en un sondeo cuya ficha técnica admitía un margen de error de ±2.8 puntos porcentuales, con una confianza de 95%. Si se lee esta información superficialmente, parece que la diferencia entre las intenciones de voto era significativamente distinta de 0, dado que 3.5 es mayor que 2.8. Sin embargo, con un análisis algo más detallado veremos que no es éste el caso: con los datos de la encuesta que publicaba El País, una diferencia en intención de voto del 3.5 puntos no es significativa.

El error muestral dado por la ficha técnica se refiere a la estimación de una sola proporción. Sin embargo, en el caso de sondeos electorales, no se estima una sola proporción, sino una colección de proporciones $(p_1, \ldots, p_K)$, cada una de las cuales corresponde a la proporción de personas que votarán a cada uno de los K partidos que concurren a las elecciones en una circunscripción. Se ha de ser consciente que el margen de error $\pm L$ es el que corresponde a intervalos de confianza para cualquiera de esas proporciones por separado. Si, por ejemplo, deseamos dar un intervalo de confianza para la diferencia entre los partidos 1 y 2, el margen de error para la diferencia $p_1 - p_2$ ya no es $\pm L$, sino que es mayor (es más difícil estimar la diferencia entre dos cantidades que estimar cada una de ellas por separado). Concretamente, y suponiendo también aquí el escenario más desfavorable (en este caso éste se da cuando $p_1 = p_2 = 0.5$ y $p_3 = \cdots = p_K = 0$, tal y como se muestra en la sección 5), el margen de error en la estimación de una diferencia de proporciones es el doble del que se tiene en la estimación de una proporción.

Volvamos al titular periodístico con el que comenzábamos esta sección. Con una confianza del 95%, el margen de error para una diferencia de proporciones –suponiendo el escenario más desfavorable– es de $\pm 2 \times 2.8 = \pm 5.6$ y, por tanto, una diferencia de 3.5% no es significativamente distinta de 0, puede atribuirse a la variabilidad muestral.

El supuesto de que el escenario que se presentará será el peor posible ($p = 1 - p = 0.5$ en la estimación de una proporción, o $p_1 = p_2 = 0.5$ y $p_3 = \cdots = p_K = 0$ en la estimación de $p_1 - p_2$) es adecuado cuando se tiene que decidir el tamaño muestral: la muestra debe ser tan grande que incluso en el peor de los caso se tenga la precisión predeterminada.

Sin embargo, una vez se ha hecho la encuesta los datos observados suelen revelar que la situación real no es la más desfavorable de todas. Ello implica que la anchura de los intervalos de confianza se puede ajustar teniendo en cuenta la información de la muestra. Las fórmulas que permiten calcular intervalos de confianza a partir de las estimaciones de proporciones y diferencias de proporciones son bien conocidas (ver, por ejemplo, Peña 1995, capítulo 4.6). A pesar de ello, su uso es prácticamente nulo en la presentación que la prensa hace de los resultados de los sondeos electorales.

A modo de ejemplo, si se usan los datos de la encuesta publicada por El País para construir un intervalo de confianza de la diferencia de votos que corresponderían a Pujol y a Maragall, se llega a que este intervalo es de $3.5\% \pm 3.9\%$. Es decir, la precisión de este intervalo no es de 5.6% (como lo sería en el caso más desfavorable de que Pujol y Maragall se repartiesen los votantes al 50%) sino que está en torno al 3.9%. En cualquier caso la diferencia de 3.5 puntos observada en la encuesta sigue sin ser estadísticamente significativa.

Sería útil que cuando se use una cifra extraída de una encuesta, ésta no apareciese sola y desamparada, sino siempre acompañada de alguna indicación sobre su precisión. Los lectores se acostumbrarían rápidamente a leer frases como *Pujol aventaja a Maragall en 3.5 (± 3.9) puntos porcentuales*, del mismo modo que ahora les es familiar ver en las fichas técnicas que el margen de error correspondientes a un nivel de confianza del 95% es de $\pm 2.8\%$.

4.2 Las horquillas de escaños

Cuando se publican los resultados de un sondeo electoral es habitual que se muestre la configuración del parlamento que corresponde a las estimaciones de las proporciones de votos estimadas para cada partido. Del mismo modo que, al estimar una proporción de votos, se ofrece un intervalo de confianza (la proporción estimada más/menos el margen de error en la estimación), cuando se estima la cantidad de escaños que corresponden a un determinado partido usualmente no se da únicamente la estimación de esa cantidad (que sería la suma de escaños que correspondería a ese partido en cada una de las circunscripciones electorales) sino que se le asigna una *horquilla de escaños*: dos números naturales entre los que previsiblemente, según el sondeo, estará el verdadero número de escaños que obtendrá ese partido finalmente en la elecciones.

La publicación de las horquillas de escaños es más informativa que el mero listado de los valores centrales de esas horquillas y por ello hemos de mostrarnos satisfechos con dicha publicación. Sin embargo, nuestra alegría no puede ser completa debido a la falta total de información sobre cómo se construyen dichas horquillas o sobre cómo deben ser interpretadas. Las fichas técnicas que se publican junto a cada sondeo ignoran por completo estos importantes aspectos.

Ninguna encuesta indica cómo se calculan las horquillas de escaños, es decir, cómo llegan desde las estimaciones de las proporciones de votos (con sus márgenes de error) a la asignación de escaños expresada en forma de horquilla o intervalo. Se supone que se aplica la ley d'Hondt circunscripción por circunscripción, pero no queda claro si se aplica esta ley a las proporciones estimadas, a las proporciones estimadas más/menos el margen de error, o a qué combinación concreta de ellas.

En ningún caso se indica la fiabilidad de las horquillas publicadas. Dicha fiabilidad se debería medir por un porcentaje de confianza, al igual que se hace en la estimación por intervalos. Al inicio de esta sección comentábamos qué se entiende por *confianza* de un intervalo: la proporción de veces que, al aplicar la misma técnica que ha producido ese intervalo, los sucesivos intervalos obtenidos contienen el verdadero valor del parámetro estimado. De modo análogo puede hablarse de la *confianza de una horquilla de escaños*: si un método para determinar horquillas de escaños tiene una confianza de, por ejemplo, el 90%, debe entenderse que las horquillas de escaños incluirían las verdaderas asignaciones de escaños en, aproximadamente, 90 de cada 100 sondeos a cuyos resultados se les aplicase dicha técnica.

No hay modelos probabilísticos sencillos que permitan definir horquillas de escaños con una confianza determinada a partir de los datos muestrales de intención de voto. Ello es debido en gran medida a que la ley d'Hondt asigna los escaños a las configuraciones de votos de forma discontinua, como hemos visto en la sección 3. Sin embargo sí es posible realizar simulaciones en el ordenador y a partir de ellas aproximar la confianza de una horquilla de escaños dada, así como elegir la horquilla más estrecha de todas aquellas que tienen al menos una cierta confianza.

Para entender cómo puede el ordenador ayudarnos a determinar horquillas de una determinada confianza (o a hallar la confianza de una horquilla dada) expondremos un problema análogo referido a la estimación de una proporción. Supongamos que queremos estimar la probabilidad p de que al lanzar una moneda equilibrada se obtenga cara (por supuesto sabemos que $p = 1/2$ y no necesitaríamos estimar ese valor, pero el ejemplo nos puede ayudar a entender situaciones más complejas). Supongamos además que queremos hacerlo tomando una muestra de lanzamientos de la moneda de tamaño $n = 50$. Llamemos $\hat{p}_{50}$ a la proporción muestral de caras en esos 50 lanzamientos. Para determinar un intervalo de confianza de (por ejemplo) el 90% para p, podemos echar mano de la aproximación de la distribución binomial por la normal. Sin embargo éste no es el único camino. Una estrategia alternativa es la siguiente. Podemos repetir tantas veces como queramos la serie de 50 lanzamientos y anotar los valores obtenidos de la proporción estimada, $\hat{p}_{50}^{(1)}, \ldots, \hat{p}_{50}^{(S)}$, donde S es el número de repeticiones del experimento. Si por ejemplo $S = 1000$ y suponemos que las distancias $d_j = |\hat{p}_{50}^{(j)} - p|$ están ordenadas de menor a mayor, se tiene que las distancias entre el estimador y el verdadero valor del parámetro serán menor que $d_{900} = |\hat{p}_{50}^{(900)} - p|$ en el 90% de los casos (aproximadamente), de donde se sigue que $(\hat{p}_{50} \mp d_{900})$ será un intervalo de confianza 90% para p. Las S series de n lanzamientos de la moneda las podríamos haber simulado con un ordenador, haciendo más cómoda la tarea.

En el ejemplo anterior había una pequeña trampa: el valor p era conocido y eso nos permitía simular datos con el ordenador que eran equivalentes a haber lanzado realmente la moneda. Pero en la realidad la proporción p no se conoce y por eso precisamente queremos estimarla. La realidad es como si sólo conociésemos los $n = 50$ resultados obtenidos al lanzar una moneda trucada (con probabilidad de cara desconocida e igual a p) y esta moneda se nos hubiese extraviado de forma que no podemos volver a lanzarla para obtener los valores $\hat{p}_{50}^{(j)}$. Aún así es posible pedir al ordenador que simule 1000 veces 50 lanzamientos de una moneda trucada con probabilidad $\hat{p}_{50}$ (la estimación de p hecha a partir de los 50 primros lanzamientos), anotar las proporciones muestrales $\hat{p}_{50}^{*(j)}$ en cada serie de lanzamientos y las distancias $d_j^* = |\hat{p}_{50}^{*(j)} - \hat{p}_{50}|$. A partir de esas distancias ordenadas, construimos el siguiente intervalo: $(\hat{p}_{50} \mp d_{900}^*)$. Pues bien, el intervalo así obtenido también tiene confianza aproximada del 90%. Obsérvese que este procedimiento no requiere ninguna información desconocida. Sólo es necesario disponer de un simulador de números aleatorios. Esta técnica basada en simulación recibe el nombre de *bootstrap paramétrico*. Véase Efron y Tibshirani (1993) para una exposición detallada sobre éste tema.

Hemos utilizado una técnica de simulación análoga a la aquí descrita para reproducir horquillas de escaños que provienen de unas determinadas proporciones de votos estimadas en cada circunscripción. También es posible estimar la confianza de una horquilla de escaños dada: es la proporción de parlamentos simulados en los que la asignación de escaños a un determinado partido está dentro de dicha horquilla. Éste es el método que hemos usado para evaluar la confianza de las horquillas que publicaron La Vanguardia, El Periódico y El País las semanas previas a las elecciones al Parlament'99. En la tabla 1 se muestran los resultados obtenidos.

	La Vanguardia		El Periódico		El País		Resultado
Tamaño muestral	n =2000		n =3643		n =2000		
	Horq.	Conf.	Horq.	Conf.	Horq.	Conf.	
CiU	56-58	56%	57-63	98%	58-60	54%	56
PSC-CC	51-52	43%	40-46	96%	48-50	57%	52
PP	13	60%	14-15	71%	13-14	60%	12
IC-V	4	48%	5-6	76%	3	52%	3
ERC	9-10	64%	12-15	96%	10	33%	12
EUA	0	99%	0	92%	0	99%	0

Tabla 1: Elecciones al Parlament de Catalunya, 17 de octubre 1999: Confianza estimada para las horquillas de escaños publicadas por diversos medios de comunicación el día 10 de octubre.
Fuente: Elaboración propia.

Como puede observarse en la tabla 1, hay gran disparidad en los criterios seguidos en las distintas encuestas publicadas. Por ejemplo, las horquillas publicadas por El País tenían una confianza aproximada del 50% para los escaños correspondientes a cada partido político, mientras que en la encuesta publicada por El Periódico las horquillas tenían una confianza de más del 95%. El aumento de la confianza se hizo a costa de ofrecer horquillas mucho más anchas que las publicadas por El País.

Para reproducir el análisis de la confianza de las horquillas de escaños que hemos presentado, la única información necesaria es el porcentaje de votos estimado para cada partido en cada circunscripción electoral. Este dato siempre está en manos de las empresas que elaboran los sondeos, por lo que éstas están en condiciones de añadir la información sobre la confianza de sus horquillas de escaños. En las elecciones al Parlament'99 también fueron publicados esos datos, pues ahí las circunscripciones son sólo cuatro. En las elecciones al congreso, el gran número de distritos electorales hace que no sea frecuente la publicación de la estimación de la intención de voto provincia por provincia. Concretamente, en vísperas de las elección al Congreso'00 de los sondeos publicados en la prensa, el de El Mundo fue el único que proporcionaba estimaciones de porcentajes de votos por provincia. Fue publicado el 5/3/00. El tamaño muestral es 12000. Los resultados se muestran en la tabla 2.

5 El tamaño muestral necesario para la predicción de diferencias o de escaños

En los sondeos electorales es habitual que se elija el tamaño muestral siguiendo la regla que se usa en el caso de querer estimar una proporción

Partido	Horquilla	Confianza	Resultado
PP	164-170	67%	183
PSOE	137-143	75%	125
CiU	16	38%	15
IU	9-11	63%	8
PNV	6-7	89%	7
CC	4-5	71%	4
BNG	3-5	95%	3
PA	0-1	93%	1
ERC	1	46%	1
IC-V	0	100%	1
EA	1	96%	1
CHA	0	95%	1

Tabla 2: Elecciones Generales, marzo de 2000: Confianza estimada para las horquillas de escaños publicadas por El Mundo.
Fuente: Elaboración propia.

poblacional p desconocida (por ejemplo, la proporción de personas a favor de una propuesta gubernamental). En ese caso se elige el tamaño muestral n de forma que los intervalos de confianza $(1-\alpha)$ para la proporción p centrados en la proporción muestral $\hat{p}$ tengan una anchura inferior a un margen de error $\pm L$ predeterminado (expresado éste en tanto por 1). La anchura de los intervalos de confianza depende del valor desconocido p: es más fácil estimar p si es un valor cercano a 1 o a 0 (por ejemplo, si p es la proporción de población a favor de reducir los impuestos que graban los combustibles) que si es un valor próximo al 50% (por ejemplo, si se quiere estimar la proporción de población que valora más la enseñanza pública que la privada). Como la verdadera proporción es desconocida antes de hacer el muestreo, se determina el tamaño muestral necesario para garantizar el margen de error $\pm L$ incluso si se está en el peor de los casos posibles, es decir, si $p = 0.5$ (o $p = 50\%$).

En definitiva, el tamaño muestral n necesario para garantizar un margen de error $\pm L$ en los intervalos de confianza 95%, suponiendo el caso más desfavorable ($p = 1-p = 0.5$) es [4] $n = 1/L^2$. Así, por ejemplo, si se desea que el margen de error sea sólo de ± 3 puntos porcentuales ($L = 3/100 = 0.03$) se necesita un tamaño muestral de 1112 personas. Para un margen de error de $\pm 5\%$ basta tomar $n = 400$. Obsérvese que si se desea reducir el margen de error a la mitad hay que cuadriplicar el tamaño muestral.

[4] Estos cálculos se basan en la aproximación de la distribución binomial por la distribución normal (véase. por ejemplo, Peña 1995, capítulo 4.6). Además se ha aproximado por 2 el cuantil 0.95 de la normal estándar, cuyo valor es 1.96. Esta aproximación permite escribir la relación entre n y L de forma más simple.

5.1 Predicción correcta de diferencias

A menudo es más interesante poder estimar con precisión la diferencia entre las proporciones de votantes de dos partidos o, más en general, entre dos coaliciones de partidos.

No es difícil deducir, como mostramos en el apéndice B, que si deseamos estimar la diferencia entre las proporciones de votos a dos partidos con igual margen de error el tamaño muestral necesario es aproximadamente cuatro veces mayor, $n \approx 4/L^2$. Esto significa que para poder afirmar que tal partido le saca 5 puntos a su competidor, con un margen de error de $\pm 3\%$ necesitaremos una muestra de 4448 votantes.

5.2 Predicción correcta de escaños

La predicción correcta de la asignación de escaños es, sin duda, uno de los principales objetivos de un sondeo electoral. Desarrollaremos una regla para determinar el tamaño muestral n cuando se quiere tener una probabilidad $(1-\alpha)$ de asignar correctamente los escaños en juego.

En el apéndice A se detalla el mecanismo de asignación de escaños basado en la ley d'Hondt. Si K partidos se deben repartir N escaños y las proporciones de votos que corresponden a cada partido son $(p_1, \ldots, p_K)$, se forman los cocientes de Hondt, se ordenan de mayor a menor y se asignan los N escaños a los partidos a los que les corresponden los N mayores cocientes. Así pues, la decisión de si es al partido A o al B aquel al que le corresponde un escaño determinado se basa en el signo de una diferencia de la forma

$$\frac{p_A}{i} - \frac{p_B}{j}.$$

En el supuesto de que sólo faltase un escaño por asignar, que A y B ya hubiesen empleado sus primeros $(i-1)$ y $(j-1)$ cocientes, respectivamente, y que los restantes cocientes aún no empleados por ningún partido fuesen menores que el más pequeño de los cocientes p_A/i y p_B/j, se tendría que el último escaño se asignaría al partido A si $(p_A/i) - (p_B/j) \geq 0$ y se asignaría a B en caso contrario.

Por lo tanto, para garantizar que la asignación de escaños se hace correctamente hay que asegurar que los signos de las diferencias de cocientes $(p_A/i) - (p_B/j)$ se estiman bien, al menos con una probabilidad alta.

Según razonamos en el apéndice B, para asignar correctamente (nivel de confianza 95%) el escaño en disputa es necesario un tamaño muestral mínimo de

$$n = 4\frac{j^2 p_A + i^2 p_B}{(j p_A - i p_B)^2} \qquad (1)$$

Esto es válido para cualquier diferencia entre cocientes $(p_A/i) - (p_B/j)$. Sin embargo sólo algunas de esas diferencias requieren ser estimadas con alta precisión: aquellas diferencias de cuyo signo depende la asignación total de escaños.

Supongamos, por ejemplo, que el número de escaños es $N = 4$, que hay $K = 2$ partidos y que $p_B = (p_A/2) + \varepsilon$, donde ε es un número positivo suficientemente pequeño: podemos pensar en $p_A = .66$ y $p_B = .34$. Los cocientes ordenados serán entonces

$$\frac{p_A}{1} = .66,\ \frac{p_B}{1} = \frac{p_A}{2} + \varepsilon = .34,\ \frac{p_A}{2} = .33,$$

$$\frac{p_A}{3} = .22,\ \frac{p_B}{2} = \frac{p_A}{4} + \frac{\varepsilon}{2} = .17,\ \frac{p_A}{4} = .165, \ldots$$

Por lo tanto, los cuatro escaños corresponden por este orden a A, B, A y A. La pequeña diferencia entre el primer cociente de B y el segundo de A no es un problema en este caso: si la estimación de las proporciones no es muy fina, puede que se llegue a que el segundo cociente de A es mayor que el primero de B (por ejemplo, se podrían obtener las estimaciones .70 y .30), pero incluso en ese caso se dará una asignación de escaños global equivalente: A, A, B y A.

Las diferencias que hay que estimar con precisión son aquellas en las que al mayor de los cocientes le corresponde un escaño mientras que al menor de ellos no le corresponde. La menor de estas diferencias es la diferencia entre el menor de los cocientes con escaño y el mayor de los que no lo tienen. En nuestro ejemplo, la diferencia que se ha de estimar bien es

$$\frac{p_A}{3} - \frac{p_B}{2} = .22 - .17 = .05.$$

El tamaño muestral requerido para una confianza del 95% será

$$n^* = 4\frac{2^2 \cdot .66 + 3^2 \cdot .34}{(2 \cdot .66 - 3 \cdot .34)^2} \approx 253.$$

Supongamos ahora que en el ejemplo anterior se disputasen 5 escaños. Entonces la diferencia importante sería

$$\frac{p_A}{4} - \frac{p_B}{2} = .165 - .17 = -.005,$$

lo cual obliga a tomar un tamaño muestral

$$n^* = 4\frac{2^2 \cdot .66 + 4^2 \cdot .34}{(2 \cdot .66 - 4 \cdot .34)^2} \approx 20200.$$

Es decir, para garantizar que con una probabilidad del 95% se asignará bien el quinto escaño es necesario tomar una muestra de 20200 personas. La muestra necesaria para asignar correctamente el quinto escaño tiene tamaño unas 80 veces mayor que la que precisábamos para asignar bien el cuarto.

Obsérvese que la expresión del tamaño muestral n depende de las probabilidades desconocidas p_A y p_B. Para que esta fórmula pueda usarse en la determinación de n antes de realizar el sondeo se precisa algún conocimiento sobre los valores de p_A y p_B, que puede proceder de un sondeo piloto o de datos históricos. Por ejemplo, n puede calcularse usando los valores de las proporciones de votos obtenidas por cada partido en las elecciones anteriores. Esto daría una regla para la afijación muestral por provincias digna de ser estudiada.

Es posible dar una regla más tosca para la elección del tamaño muestral que puede usarse sin estimaciones previas de las proporciones verdaderas. En el apéndice B.2 se muestra que el tamaño muestral necesario es a los sumo

$$n \approx \frac{4}{L^2}, \qquad \text{donde } L = |(j/i)p_A - p_B|.$$

Esta regla simple es muy similar a la que hemos dado para la estimación de diferencias. El valor $(j/i)p_A$ puede interpretarse como una corrección de la proporción p_A para hacerla comparable con p_B. Así, L sería la máxima diferencia entre proporciones corregidas que estamos dispuestos a aceptar. Podría usarse el valor $n = 4/L^2$ con la seguridad de que con una probailidad del 95% ésta diferencia L no sería superada. Como valores de L podrían usarse, por ejemplo, 0.1 o 0.05 (véase la tabla con la que concluye el apéndice B.2). El inconveniente que presenta el uso de esta fórmula genérica es que los tamaños muestrales a que da lugar son considerablemente más altos que los obtenidos a partir de una estimación previa de las proporciones desconocidas y la aplicación de la fórmula (1).

6 Conclusiones

En este artículo se han propuesto métodos gráficos para visualizar conjuntamente los resultados de diversos sondeos electorales, se han estudiado algunos de sus aspectos estadísticos (entre ellos, el significado de las horquillas de escaños) y los efectos que sobre los resultados de un sondeo tienen las peculiaridades de la ley d'Hondt. En particular, hemos mostrado que esta ley introduce en los sondeos un sesgo importante en la estimación del parlamento final. Conviene desarrollar métodos de corrección de este sesgo, bien sea por métodos de MonteCarlo que puedan estimarlo a partir de los datos, como se explica al final de la sección 4, bien mejorando la afijación muestral que no se debe realizar en proporción al tamaño censal de las circunscripciones sino de la dificultad de la estimación de los escaños respectivos y ésta se puede estimar a partir de las características que se deriven de un sondeo piloto o de resultados previos.

Referencias

Bernardo, José M. (1984). Monitoring the 1982 spanish socialist victory: A bayesian analysis. *JASA*, **79**, 510–515.

Cuadras, C. M. (1996). *Métodos de análisis multivariante*. EUB.

Efron, B. y R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Peña, D. (1995). *Estadística: Modelos y Métodos*, Volumen 1: Fundamentos. Alianza Universidad, Madrid. Segunda edición revisada.

A Expresión matemática de la ley d'Hondt como función

Detallamos primero la forma en que debe aplicarse la ley d'Hondt. La tabla de la figura 6 muestra el cálculo para el caso de cuatro partidos que se disputan seis escaños.

	Partidos			
	A	B	C	D
votos	f_1	f_2	f_3	f_4
$j=1$	$\boxed{f_1}$	$\boxed{f_2}$	f_3	$\boxed{f_4}$
$j=2$	$\boxed{f_1/2}$	$\boxed{f_2/2}$	$f_3/2$	$f_4/2$
$j=3$	$\boxed{f_1/3}$	$f_2/3$	$f_3/3$	$f_4/3$
$j=4$	$f_1/4$	$f_2/4$	$f_3/4$	$f_4/4$
$j=5$	$f_1/5$	$f_2/5$	$f_3/5$	$f_4/5$
$j=6$	$f_1/6$	$f_2/6$	$f_3/6$	$f_4/6$

Figura 6: Tabla para la aplicación de la regla d'Hondt en un caso de cuatro partidos que se disputan seis escaños. Bajo los nombres de los partidos figuran el número de votos obtenidos f_i. Para cada j desde 1 hasta el número de escaños se forma el cociente f_i/j. Se marcan los seis cocientes mayores y en este caso resultarían tres escaños para el partido A, dos para el B, ninguno para el C y un sólo escaño para el partido D.

Veamos ahora cómo se puede generalizar el estudio de algunos casos que hemos mostrado anteriormente.

Sean K partidos que se disputan N escaños. Sea δ el umbral de proporción por debajo del cual no se puede obtener ningún escaño. Sean

$(p_1, p_2, \ldots, p_K)$ las proporciones de votos respectivas, de modo que

$$0 \leq p_i \leq 1, (i = 1, \ldots, K), \qquad \sum_{i=1}^{K} p_i = 1.$$

Sean $q_{i,j}$, $(i = 1, \ldots, K,\ j = 1, \ldots, N)$ los llamados cocientes d'Hondt definidos por:

$$\begin{array}{lll} \text{Si } p_i < \delta, & q_{i,j} = 0 & \text{para } j = 1, \ldots, N \\ \text{Si } p_i \geq \delta, & q_{i,j} = p_i/j, & \text{para } j = 1, \ldots, N \end{array}$$

La regla d'Hondt asigna un escaño para cada uno de los N cocientes mayores, después de ordenar los cocientes $q_{i,j}$, $i = 1, \ldots, K$, $j = 1, \ldots, N$ de mayor a menor. En caso de un improbable empate, asignaría el escaño al partido con p_i mayor.

Esta regla se puede caracterizar como una función H del simplex $0 \leq \sum_{i=1}^{K} f_i \leq 1 \subset \mathcal{R}^K$ en $\mathcal{N}^K$ con

$$\begin{aligned} & S_d(f_1, \ldots, f_K) = (m_1, \ldots, m_K) \Longleftrightarrow \\ & \quad \forall i, j \in \{1, \ldots, K\}, i \neq j, \quad m_i = 0 \text{ o } \frac{f_i}{i} > \frac{f_j}{j+1} \end{aligned} \tag{2}$$

También se puede ver que, con la notación anterior,

$$\begin{aligned} & S_d(f_1, \ldots, f_K) = (m_1, \ldots, m_K) \Longleftrightarrow \\ & \quad \forall i = 1, \ldots K, \quad m_i = \max\{j = 1, \ldots, M | Q(i,j) > KN - N\} \\ & \quad \text{donde } Q(i,j) = \#\{q_{k,l} < q_{i,j} : k = 1, \ldots, K, l = 1, \ldots, M\} \end{aligned} \tag{3}$$

puesto que el último escaño asignado al partido i debe dejar por debajo al menos tantos cocientes como $KN - N$. Nótese que $Q(i,j)$ es el número de cocientes por debajo de $q_{i,j}$.

De las $K(K-1)$ desigualdades que aparecen como máximo en (2), algunas pueden ser redundantes, pero en cualquier caso resulta que la función H es discontinua, siendo constante en poliedros convexos de $\mathcal{R}^K$ (que son regiones delimitadas por hiperplanos en el simplex). En la figura 7 se pueden ver los poliedros (polígonos en este caso) con H constante para $K = 3$ partidos que se disputan $N = 6$ escaños. Obsérvese que las celdas centrales son hexagonales, lo que significa que las seis desigualdades de (2) están activas, mientras que las celdas adyacentes a los lados del triángulo son cuadriláteras.

B Estimación simultánea de proporciones y diferencias

Según hemos visto en la sección 5, en la estimación de una proporción con nivel de confianza 95%, el tamaño muestral mínimo debe ser $n \approx 1/L^2$, donde L es el margen de error máximo aceptable.

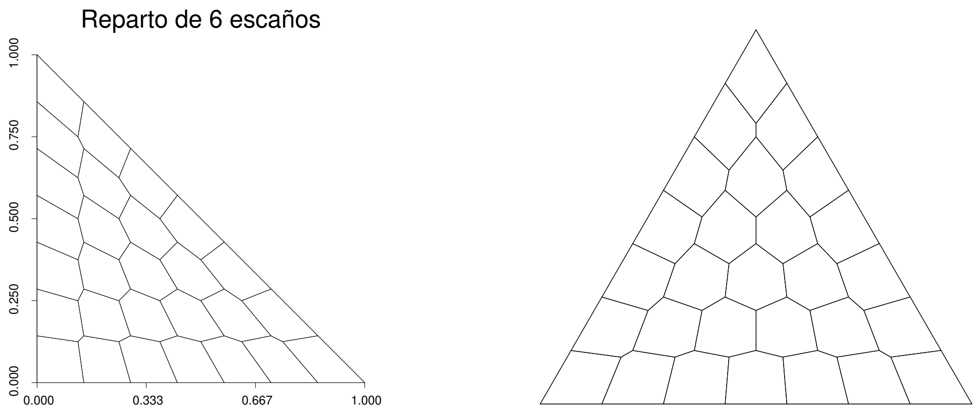

Figura 7: Reparto de seis escaños según las proporciones de voto para 3 partidos. Cada punto del interior del triángulo corresponde a un reparto de votos, cada celda delimita aquellos repartos de votos que dan lugar a una idéntica distribución de escaños. A la izquierda se usan coordenadas cartesianas con las proporciones de dos de los partidos en los ejes. A la derecha, coordenadas triangulares en que la distancia a los lados del triángulo son proporcionales a las proporciones de voto de los partidos.
Fuente: Elaboración propia.

Pongamos ahora que no se estima una sola proporción, sino la colección de proporciones $(p_1, \ldots, p_K)$ que corresponden a cada uno de los K partidos existentes. Típicamente, el resultado de un sondeo será una colección de estimaciones de esas proporciones: $(\hat{p}_1, \ldots, \hat{p}_K)$. En esta sección nos ocuparemos del problema de la elección del tamaño muestral teniendo en cuenta que son K las proporciones que se estiman simultáneamente.

Si actuamos por analogía con el caso de estimación de una proporción, el tamaño muestral n se debería elegir para garantizar que el error de estimación sea menor que una cantidad fijada L, con una cierta confianza $(1-\alpha)$. Ahora bien, el término *error de estimación* tenía un significado claro en la estimación de una proporción (es la distancia entre el estimador $\hat{p}$ y el verdadero valor de p: $|\hat{p}-p|$) mientras que no es tan sencillo definir qué se entiende por *error de estimación* cuando se estiman simultáneamente K proporciones.

Hay diversas formas de medir la distancia entre las estimaciones $\hat{P} = (\hat{p}_1, \ldots, \hat{p}_K)$ y las proporciones reales $P = (p_1, \ldots, p_K)$: podrían calcularse las K distancias $|\hat{p}_i - p_i|$ y quedarse con la máxima de ellas, o hacer un promedio; o calcular la distancia entre $\hat{P}$ y P como elementos de un espacio euclídeo de dimensión K; o calcular versiones ponderadas de la distancia euclídea más indicadas en este caso (como la distancia de Mahalanobish o la distancia χ^2; véase Peña 1995, apéndice 3G, o Cuadras 1996, para las definiciones concretas de estas distancias). En general, estas distancias son difícilmente interpretables en términos intuitivos (por ejemplo, la distancia euclídea es la raíz cuadrada de la suma de los cuadrados de las distancias que separan cada estimación $\hat{p}_i$ de la proporción p_i).

Hemos usado aquí dos criterios para medir distancias entre las proporciones reales y sus estimaciones: por una parte, la diferencia de proporciones de votos asignadas a dos coaliciones de partidos, y por otra, la discrepancia entre las asignaciones de escaños a que dan lugar las proporciones de votos.

B.1 Diferencias entre coaliciones

A menudo se desea estimar con precisión la diferencia de las proporciones de votantes de dos partidos o, más en general, de dos coaliciones de partidos. Veamos cuál es el peor escenario posible cuando se estima la diferencia entre las proporciones de dos coaliciones de partidos.

Dos coaliciones de partidos pueden representarse mediante un vector $a = (a_1, \ldots, a_K)$ de longitud K igual al número de partidos cuyos elementos sean 1 o -1: los partidos a los que corresponde un 1 son de una coalición y aquellos a los que les corresponde un -1 forman la otra. La diferencia entre la suma de proporciones de votos de ambas coaliciones es

$$d(a,p) = \sum_{i=1}^{K} a_i p_i.$$

La estimación de esta diferencia a partir de los datos de la muestra es

$$d(a,\hat{p}) = \sum_{i=1}^{K} a_i \hat{p}_i.$$

Buscar el reparto de votos entre partidos que hace más difícil estimar la diferencia entre las dos coaliciones equivale a buscar las proporciones $(p_1, \ldots, p_K)$ que hacen máxima la varianza de $d(a, \hat{p})$. Según las propiedades de la distribución multinomial (ver, por ejemplo, Peña 1995), la varianza de este estimador es

$$V(d(a,\hat{p})) = \frac{1}{n}\left[\sum_{i^1}^{K} a_i^2 p_i - \left(\sum_{i=1}^{K} a_i p_i\right)^2\right].$$

Como los elementos a_i son 1 o -1, su cuadrado siempre vale 1 y, por tanto,

$$\sum_{i^1}^{K} a_i^2 p_i = \sum_{i=1}^{K} p_i = 1$$

lo cual implica que

$$V(d(a,\hat{p})) = \frac{1}{n}\left[1 - \left(\sum_{i=1}^{K} a_i p_i\right)^2\right].$$

Así, buscar la situación más desfavorable (la que da lugar a $V(d(a,\hat{p}))$ máxima) equivale a buscar las proporciones que hacen mínimo el valor $\left(\sum_{i=1}^{K} a_i p_i\right)^2$. El mínimo valor que puede tomar ese cuadrado es 0. Ese es precisamente el valor que toma si las dos coaliciones suman una proporción de votos del 50% cada una. En particular, esto ocurre si sólo dos partidos se reparten los votos a partes iguales: $p_1 = p_2 = 0.5$, $p_3 = \cdots = p_K = 0$, $d(a,p) = p_1 - p_2$. Así, en el peor de los casos la varianza del estimador de $d(a,p)$ coincide con la varianza del estimador de $(p_2 - p_1)$ en el caso más desfavorable y tiene un valor de $1/n$. Eso da lugar a que los intervalos de confianza $(1-\alpha)$ para $(p_1 - p_2)$ sean, en el peor de los casos, de la forma

$$\left((\hat{p}_1 - \hat{p}_2) \pm \frac{z_{\alpha/2}}{\sqrt{n}}\right),$$

donde z_α es el cuantil $(1-\alpha)$ de la distribución normal estándar (por ejemplo, si $(1-\alpha) = 0.95$, entonces $z_{\alpha/2} = 1.96 \approx 2$). Sea cual sea la confianza $(1-\alpha)$ deseada, estos intervalos son el doble de anchos que los intervalos con la misma confianza construidos para estimar una única proporción p.

Si se desea estimar la diferencia entre dos coaliciones con un margen de error inferior a L, con una confianza de $(1-\alpha)$, incluso si los votos se reparten de la peor forma posible, el tamaño muestral necesario es

$$n = \frac{z_{\alpha/2}^2}{L^2}$$

que es cuatro veces el tamaño necesario para estimar una proporción con idéntica precisión. Si $\alpha = 0.05$, entonces $n \approx 4/L^2$.

B.2 Predicción de escaños

Hemos visto en la sección 5 que para la correcta asignación de escaños interesa estimar diferencias del tipo $(p_A/i) - (p_B/j)$. Desarrollamos ahora la regla que allí se ha enunciado.

El estimador natural de esa diferencia se construye a partir de la proporciones estimadas de votos para los dos partidos y es

$$\frac{\hat{p}_A}{i} - \frac{\hat{p}_B}{j}.$$

Nos planteamos elegir el tamaño muestral n necesario para asegurar que, con probabilidad mayor o igual que $(1-\alpha)$, los signos de las diferencias

$$\frac{p_A}{i} - \frac{p_B}{j} \text{ y } \frac{\hat{p}_A}{i} - \frac{\hat{p}_B}{j}$$

coincidan.

Por las propiedades de la distribución binomial y por el Teorema Central del Límite, se tiene que si el tamaño muestral es suficientemente grande entonces

$$\frac{\hat{p}_A}{i} - \frac{\hat{p}_B}{j} \sim_A N\left(\frac{p_A}{i} - \frac{p_B}{j}, \sigma^2(p_A, p_B, i, j)\right),$$

donde el símbolo $\sim_A$ significa que la variable aleatoria de la izquierda tiene distribución aproximada a la que se escribe a la derecha, y la varianza de la diferencia de cocientes estimados es

$$\begin{aligned} &\sigma^2(p_A, p_B, i, j) \\ &= \frac{1}{i^2}\frac{p_A(1-p_A)}{n} + \frac{1}{j^2}\frac{p_B(1-p_B)}{n} + 2\frac{1}{ij}\frac{p_A p_B}{n} \\ &= \frac{j^2 p_A(1-p_A) + i^2 p_B(1-p_B) + 2ij p_A p_B}{ni^2 j^2}. \end{aligned} \tag{4}$$

Así, si Z es una variable aleatoria normal estándar, se tiene que

$$P\left(\frac{\hat{p}_A}{i} - \frac{\hat{p}_B}{j} > 0\right) \approx P\left(Z > \frac{-\left(\frac{p_A}{i} - \frac{p_B}{j}\right) ij\sqrt{n}}{\sqrt{j^2 p_A(1-p_A) + i^2 p_B(1-p_B) + 2ij p_A p_B}}\right),$$

y queremos que esta probabilidad sea $(1-\alpha)$. Llamemos z_α al número real tal que $P(Z > z_\alpha) = \alpha$. Por simetría de la distribución normal, $P(Z > -z_\alpha) = 1 - \alpha$. Se sigue que

$$\frac{\sqrt{n}(jp_A - ip_B)}{\sqrt{j^2 p_A(1-p_A) + i^2 p_B(1-p_B) + 2ij p_A p_B}} = z_\alpha,$$

despejando n y reordenando algunos términos,

$$n = z_\alpha^2 \left[\frac{j^2 p_A + i^2 p_B}{(jp_A - ip_B)^2} - 1\right]. \tag{5}$$

Para un nivel de confianza del 95%, $\alpha = 0.05$, valdrá la aproximación

$$n \approx 4\frac{j^2 p_A + i^2 p_B}{(jp_A - ip_B)^2}. \tag{6}$$

Analicemos la expresión de n encontrada. Cuanta mayor sea la confianza $(1-\alpha)$ deseada, mayor será el número z_α y, por lo tanto, mayor debe ser el tamaño muestral, como es lógico. Por otra parte, $jp_A - ip_B$ es pequeño (respectivamente, grande) si y solo si lo es la diferencia de los cocientes que se desea estimar, y n depende inversamente del cuadrado de esta diferencia. Así, cuanto más cerca de 0 está la diferencia que se quiere estimar, mayor debe ser el tamaño muestral empleado.

Es posible simplificar algo la expresión anterior de n. Obsérvese que si, por ejemplo, $p_A > p_B$, se tendrá que $i \geq j$, porque de lo contrario la asignación del escaño en cuestión no dependería de la diferencia $(p_A/i) - (p_B/j)$, sino de otra diferencia de cocientes en la que el denominador de p_B fuese menor que j, así $(j/i) \leq 1$ y, por tanto, $(j/i)^2 \leq (j/i) \leq 1$. Así,

$$n \approx z_\alpha^2 \frac{(j^2/i^2)p_A + p_B}{((j/i)p_A - p_B)^2} \leq \frac{z_\alpha^2}{((j/i)p_A - p_B)^2}.$$

Definimos $L = |(j/i)p_A - p_B|$, el valor absoluto de la diferencia entre la proporción mayor p_A ajustada (multiplicada por (j/i) para hacerla comparable con la proporción menor p_B) y la proporción menor p_B. Esta cantidad puede considerarse como la precisión deseada en el sondeo (o el margen de error permitido): se desea fijar n de forma que si la diferencia entre proporciones (ajustada la mayor) es mayor que esa precisión L, entonces la probabilidad de estimarla bien sea de al menos $(1 - \alpha)$. La definición de L permite expresar las diferencias entre cocientes en otra escala, en la que pueden interpretarse como diferencias entre proporciones. Además, ayuda a simplificar la expresión del tamaño muestral.

Así,

$$n \leq \frac{z_\alpha^2}{L^2}.$$

Por lo tanto, si se elige el tamaño muestral

$$n^* = \frac{z_\alpha^2}{L^2}, \tag{7}$$

se está garantizando que se cumplen los objetivos marcados. Obsérvese que esta fórmula para el tamaño muestral coincide con la que, en la sección 5, se recomienda usar en la elección del tamaño muestral cuando se estiman diferencias de proporciones.

Si $1 - \alpha = 0.95$, entonces $n^* = 4/L^2$. Despejando L en función de n^* se tiene que

$$L = \frac{2}{\sqrt{n^*}}.$$

Estas fórmulas permiten completar la tabla siguiente, cuando la confianza fijada es del 95%.

$L = \lvert(j/i)p_A - p_B\rvert$	0.20=20%	0.10=10%	0.05=5%	0.032=3.2%
n	100	400	1600	4000

C Datos técnicos de los sondeos citados

Los sondeos previos a las elecciones al Parlament de Catalunya 1999 que hemos utilizado responden a las siguientes características.

Periódico	Fecha	Empresa	Tamaño muestral
La Vanguardia	10/10/99	I. Opina	2000
El País	10/10/99	Demoscopia	2000
El Mundo	10/10/99	Sigma Dos	2000
Diario 16	10/10/99	Colpisa/Metra Seis	2000
ABC	10/10/99	IPSOS-Eco C.	2400
El Periódico	10/10/99	DYM	3643
C.I.S.	8/10/99	C.I.S.	3590

Fuente: Información publicada por la Generalitat de Catalunya en sus páginas web dedicadas al seguimiento de las elecciones al Paralament de Catalunya, 1999.

La mayoría de los sondeos utilizaron para la afijación muestral en la cuatro provincias catalanas la regla de asignar tamaño doble a la de Barcelona que a las otras tres, que tuvieron idéntica afijación.

En el caso de las elecciones generales al Congreso 2000, los sondeos utilizados han sido los siguientes.

Periódico	Fecha	Empresa	Tamaño muestral
La Vanguardia	5/3/2000	I. Opina	3000
El País	5/3/2000	Demoscopia	15000
El Mundo	5/3/2000	Sigma Dos	12000
ABC	5/3/2000		2300
El Periódico	5/3/2000	Vox Pública	15600

Fuente: Elaboración propia a partir de los datos publicados por los medios de comunicación mencionados.

Sólo el sondeo de El País ofrecía detalles sobre la afijación por provincias. Según su ficha técnica, se asignaron 100 encuestas a cada circunscripción y el resto se repartió proporcionalmente a la población censada.

Pedro Delicado. Currículum Vitae

Es licenciado en Ciencias Matemáticas (especialidades de Estadística e Investigación Operativa) en la Universidad Complutense de Madrid, y Doctor en Economía por la Universidad Carlos III de Madrid (1995). Tras unos años como profesor en la Universitat Pompeu Fabra, en estos momentos es profesor del Departamento de Estadística e Investigación Operativa de la Universitat Politècnica de Catalunya. Ha publicado diversos artículos de investigación en revistas internacionales, en campos que van desde la estadística teórica (en revistas como *Multivariate Analysis* o *Annals of the Institute of Statisitcal Mathematics*) hasta las aplicaciones en economía o medicina (en *Journal of Forecasting* o *Investigative Radiology*).

Frederic Udina. Currículum Vitae

Es Licenciado en Matemáticas (Universidad Autònoma de Barcelona) y Doctor en Matemáticas (Departamento de Estadística e Investigación Operativa, Universitat Politècnica de Catalunya). Desde hace más de diez años es profesor de matemáticas y estadística en la Universtitat Pompeu Fabra. Los campos en los que desarrolla su investigación son la estimación no paramétrica de curvas, la estadística computacional, los gráficos dinámicos, la metodología y aspectos prácticos de la docencia de las matemáticas y el uso de software para la enseñanza de las matemáticas, la probabilidad y la estadística. Ha publicado artículos en revistas como *Journal of Statistical Software*, *Statistics and Computing* o *Computational Statistics and Data Analysis*.