¿Qué tan *"normal"* fue la elección del 2006?

Jorge A. López, Ph.D.

Al calor de las intensas discusiones que se suscitaron en la comunidad científica de México e internacional después de la elección presidencial del 2006, me argumentaba un famoso colega, en referencia al estudio del Dr. Víctor Romero de la UNAM:

> "yo no estoy tan convencido de su análisis sobre el fraude... Mi impresión es que los estadísticos acostumbrados a variables físicas, no se dan cuenta que los datos electorales no son variables estocásticas, hay muchas correlaciones implícitas que no tomaron en cuenta."

Aunque esta respuesta no correspondía plenamente al estudio de Romero, que estuvo basado en un estudio temporal de los datos del PREP, me vi sin argumentos para aceptar o rebatir el comentario del colega. La verdad es que en ese momento sabía poco, o casi nada, del comportamiento que deberían tener los votos en una elección, por lo que en ese momento callé. Pero en este caso, el que calla no otorga, sino que se pone a estudiar.


Dr. Víctor Romero

Varios asuntos salen a la superficie de ese comentario: ¿son variables estocásticas los datos de los votos, o no?. Es más, ¿son variables los votos? ¿Qué tan variables son los votos? Y de serlo ¿son estocásticos o dependen de algún factor.

Lo que nos lleva al segundo ingrediente del comentario: ¿hay correlaciones que se deben tomar en cuenta? Es decir, si los votos dependen de otros factores, ¿cómo estos afectan la votación? Muchas preguntas y muy pocas respuestas.

estocástico, ca.
(Del gr. στοχαστικός, hábil en conjeturar).

1. adj. Perteneciente o relativo al azar.

2. f. *Mat.* Teoría estadística de los procesos cuya evolución en el tiempo es aleatoria, tal como la secuencia de las tiradas de un dado.

variable estocástica

1. f. *Mat.* Magnitud cuyos valores están determinados por las leyes de probabilidad, como los puntos resultantes de la tirada de un dado.

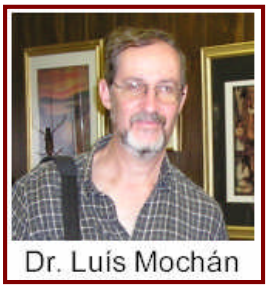
Este estudio trata de saber si los datos de la elección pasada tienen un comportamiento "*entendible*", es decir, uno que ya haya sido visto en algún otro sistema. De ser así podremos usar lo que se sabe sobre estos sistemas para identificar desviaciones de esta asumida "*normalidad*".

Debido a que los argumentos que aquí se presentarán son de índole matemática, posible *coco* de muchos lectores, se ha hecho un esfuerzo por explicar todos los cálculos de una manera pedagógica incluyendo numerosos ejemplos.

Pero, para aquellos que les da pereza seguir argumentación matemática, les resumo el trabajo y los resultados en un párrafo. Se analizaron los datos de la elección presidencial desde un punto de vista estadístico. Dividiendo los datos por estados y limitando el estudio a los tres partidos mayores, se observó que los datos tienen una fuerte tendencia a seguir una distribución gaussiana, comportamiento común de sistemas diversos. Esto abre la puerta al uso de tecnología conocida para establecer una medida de la *normalidad* de la elección, y –por el contrario- identificar desviaciones de este patrón. Al aplicar tales técnicas, se detectaron fuertes desviaciones en varios estados, y se encontró que los resultados de cientos de casillas tienen probabilidades de existencia menores a 1 en 10

millones; dado que hubo nada más 130 mil casillas la probabilidad de que se den esos resultados en una elección "*normal*" es prácticamente cero.

Los datos


Dr. Luís Mochán

Los análisis iniciales [Mochán, Romero, López...] se hicieron durante la tormenta que suscitó la publicación de los datos del PREP. Ahora ya con la calma que siguió a esa tempestad, es posible analizar los datos finales. El archivo con todos los datos del conteo distrital para la elección de presidente se puede copiar del sitio de Luis Mochán:
http://em.fis.unam.mx/~mochan/elecciones/Computos2006-Presidente.zip

el cual fue copiado directamente de la base de datos del IFE:
http://www.ife.org.mx/Computos2006/bd_computos06.htm
que, desgraciadamente, no está más en operación.

Estos datos contienen un poco más de 130,000 líneas con la siguiente información y en el siguiente formato:

```
1|1|0|1|B|0|6||120|0|0|76|76|1|53|4|18|0|1|0|||05/07/2006 23:22:10
1|1|338|1|B|0|1|1|689|12|5|341|353|1|202|45|65|12|12|2|1|1|05/07/2006 08:17:05
1|1|338|1|C|0|1|1|689|7|11|291|298|2|175|38|56|3|8|2|1|1|05/07/2006 08:22:24
. . .
32|4|1708|1|B|0|1|1|273|1|0|165|166|515|6|75|82|0|2|54|1|1|05/07/2006 18:35:01
32|4|1709|1|B|0|1|1|277|9|1|175|184|516|13|51|108|0|2|54|1|1|05/07/2006 18:35:17
32|4|1710|1|B|0|1|1|485|9|8|183|192|517|23|63|85|2|2|54|1|1|05/07/2006 18:35:36
```

Cada línea corresponde a una casilla y cada número significa, en orden:

```
ID_ESTADO, DISTRITO, SECCION, ID_CASILLA, TIPO_CASILLA, EXT_CONTIGUA,
TIPO_CANDIDATURA, TIPO_ACTA, LISTA_NOMINAL, NO_VOTOS_NULOS, NO_VOTOS_CAN_NREG,
NO_VOTOS_VALIDOS, TOTAL_VOTOS, ORDEN, PAN, APM, PBT, NA, ASDC, MUNICIPIO,
PAQUETE_ENTREGADO, CASILLA_INSTALADA, FECHA_HORA
```

Donde los candidatos están identificados como PAN, APM (Alianza por México), PBT (Coalición Por El Bien De Todos), NA (Nueva Alianza) y ASDC (Alianza Social Demócrata Campesina). Cabe notar que veinte casillas contienen datos incompletos del número de votos por lo que tuvieron que ser eliminadas, Estas casillas son:

```
2|7|732|1|B,   12|9|348|1|B,   12|9|349|1|B,   13|3|682|1|B,   15|38|4642|1|B,
15|38|4642|1|C,  20|5|197|1|B,  20|5|1472|1|B,  20|5|1472|1|C,  20|5|1472|2|C,
20|6|132|1|B,   20|6|2045|2|E,  20|6|2046|1|B,  20|6|2046|1|C,  20|6|2046|1|E,
20|6|2047|1|E, 20|6|2048|1|E, 20|6|2048|1|B, 20|6|2048|1|C, 20|7|1191|1|E,
```

Los porcentajes

Para poder combinar y comparar resultados de casillas distintas, es conveniente trabajar, no con los votos directamente, sino con los ***porcentajes*** obtenidos por casillas. Esto se logra dividiendo los votos de un candidato por el total de votos válidos, y multiplicando por 100.

Por ejemplo, para el caso de la siguiente casilla de Zacatecas:

```
32|4|1708|1|B|0|1|1|273|1|0|165|166|515|6|75|82|0|2|54|1|1|05/07/2006 18:35:01
```

el número de votos válidos fue de 165, los votos del PAN fueron 6, la APM obtuvo 75, la Alianza PBT ganó 82, la NA no tuvo votos, y la ASDC logró 2 votos. En términos de porcentajes, esta información se transforma a:

Número total de votos: 165		
Partido	Número de votos	Porcentaje
PAN	6	$\frac{6}{165}\times 100 = 3.636363\ \%$
APM	75	$\frac{75}{165}\times 100 = 45.454545\ \%$
PBT	82	$\frac{82}{165}\times 100 = 49.696969\ \%$
NA	0	$\frac{0}{165}\times 100 = 0\ \%$
ASDC	2	$\frac{2}{165}\times 100 = 1.212121\ \%$
Total	165	100 %

Con esto, los datos de todas las casillas, grandes y chicas, se transforman a números entre cero y cien. Por ejemplo las tres casillas de Zacatecas mostradas en la página anterior (estado 32) ahora son:

% PAN	% APM	% PBT	% NA	% ASDC
3.636363636	45.45454545	49.6969697	0	1.212121212
7.428571429	29.14285714	61.71428571	0	1.142857143
12.56830601	34.42622951	46.44808743	1.092896175	1.092896175

Esta transformación se puede lograr fácilmente por medio de un programa de computadora. En el presente estudio esto se hizo por medio del programa *Excel* que usa hojas de trabajo y programación en "*Visual Basic*".

La distribución de porcentajes

El siguiente paso en nuestra búsqueda de la "*normalidad*" es estudiar estos porcentajes.

Para entender la utilidad de un estudio por porcentajes, considere el siguiente ejemplo. Se sabe que las preferencias por los candidatos varían de estado a estado. Mientras que –por ejemplo- en el norte del país se espera que el PAN tenga porcentajes altos, asimismo se piensa que este porcentaje debe disminuir en el centro y sur de la república. Así pues, si en un estado sureño donde el promedio del PAN es –digamos- del 12%, apareciera alguna casilla con –digamos- un 98% de los votos para el PAN, es de esperarse que eso levantara sospechas. ¿Pero acaso sucedería lo mismo con un porcentaje de 60%? ¿O de 30%?

Eso nos lleva a la pregunta: ¿cómo decidir si un resultado es aceptable o no? La respuesta nos la da la ***distribución de porcentajes***.

Para obtener esa distribución simplemente es necesario contar. Digamos que nos interesa saber cuantas de las dos mil y pico de casillas electorales de Zacatecas terminaron con porcentajes para la APM de entre 21% y 23%. Analizando los dos mil y pico de porcentajes de la APM, uno simplemente cuenta los que caen entre estos valores y ya. Repitiendo esto para todos los valores posibles de porcentajes se obtiene la distribución deseada.

% APM
A=33.4841629
B=22.10526316
C=21.83908046
D=25
E=20.9486166
F=23.01886792
34.30420712
40.32786885
31.29251701
16.04095563
19.34306569
26.71755725
26.47058824
37.77777778
44.40298507
43.81679389
20.08196721
16.47058824
19.18367347
17.59656652
27.22772277
29.13669065
30.73929961
24.90842491
16.73003802
21.57676349
23.64341085
23.73540856
28.40466926

Por ejemplo, consideremos los porcentajes de la tabla de la izquierda, y clasifiquémoslos en clases de 2%. Es decir, contemos los que caen del 0% → 2%, del 2% → 4%, ... hasta llegar al 98% → 100%. A manera de nombre para cada clase usaremos el valor del medio de la misma, *eg.* la del 0% → 2% será llamada la del 1%, la siguiente será la del 3%, y así hasta llegar a la del 99%.

Con esto, es fácil ver que la clasificación de los primeros seis valores de la tabla resulta en el siguiente conteo, donde las letras identifican los valores:

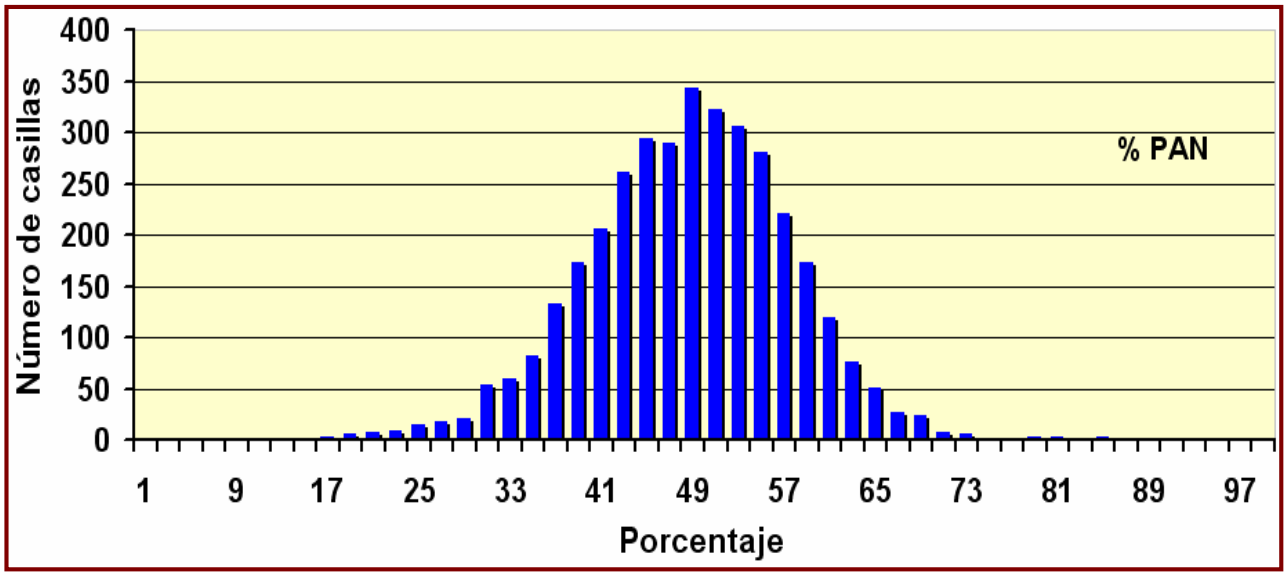
```
          E  F
          C  B  D           A
17 19 21 23 25 27 29 31 33 35 37 39 41 43 45
```

Continuando con el resto de la columna (y usando simplemente una "X" para indicar la existencia de un valor con ese porcentaje), obtenemos:

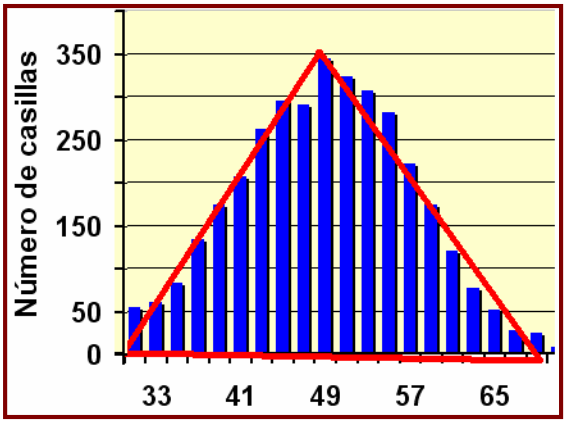
```
X     X  X
X     X  X     X
X  X  X  X  X  X  X  X
X  X  X  X  X  X  X  X  X  X  X     X  X  X
17 19 21 23 25 27 29 31 33 35 37 39 41 43 45
```

Es decir, 4 casillas quedaron en la clase del 17 %, que abarca porcentajes del 16 % → 18 %; 2 en la del 19 %; 4 en 21 %; et cetera.

Aplicando este procedimiento a la votación recibida por el PAN en todas las casillas de Baja California, se obtiene la siguiente gráfica:



Lo cual despeja la duda de cómo debe ser la distribución de votos en una elección. Esa distribución se asemeja mucho a la distribución gaussiana, que sirve para describir infinidad de procesos naturales, desde la distribución de variedades biológicas hasta la distribución de alturas de niños en primaria. Pero eso será revisado después, por ahora conviene hacer una prueba somera para comprobar que no andamos muy errados.



Para ver si los resultados corresponden a *grosso modo* con la realidad, podemos estimar el número de casillas que aparecen en la distribución. Tomando la figura de la distribución crudamente como un triángulo, su área nos dará una aproximación al número de casillas incluidas en el estudio. Dado que el área es $\frac{\text{Base} \times \text{Altura}}{2} \approx \frac{(69-31)\times 350}{2} \approx 6650$, y dado que la base está dada en unidades de 1% y como cada columna representa un 2%, el número de casillas es $\approx 6300/2 = 3325$, lo que se aproxima bien al número de casillas de Baja Califonia incluidas en este estudio: 3549; como dirían los gringos, en traducción libre, no sabremos jugar, pero al menos ya estamos en el campo correcto.

La curva gaussiana

La curva gaussiana, también conocida como curva normal, segunda ley de Laplace, ley del error, etc., fue descubierta en el siglo 18 por de Moivre, pero fueron los trabajos de Laplace, y Gauss –que relacionaron la distribución de errores en mediciones experimentales con la curva normal- los que la ubicaron en un lugar especial en el mundo de las matemáticas.

[Gauss fue honrado por el pueblo Alemán al aparecer en la última edición de billetes de 10 marcos junto con su famosa curva.]

La curva gaussiana se usa para explicar infinidad de distribuciones de mediciones. Por ejemplo, los resultados de las pruebas de aptitud de preparatoria de los EE.UU. siguen esta distribución, así como las mediciones del coeficiente intelectual (IQ), la altura de las personas, y hasta las ondas de radio que llegan a la tierra del espacio exterior, el mismo Gauss la usó para explicar distribuciones de datos astronómicos.

No se sabe porqué la curva normal describe tantos y tan distintos procesos, su uso –sin embargo— puede ser justificado teóricamente cuando existen muchos efectos independientes que se suman a cada observación; suposición ciertamente válida en el caso de las elecciones.

Por su vasta utilidad la curva gaussiana ha sido estudiada ampliamente. Se sabe, por ejemplo, que la expresión matemática que la describe es (ver ampliación del billete de 10 DM):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

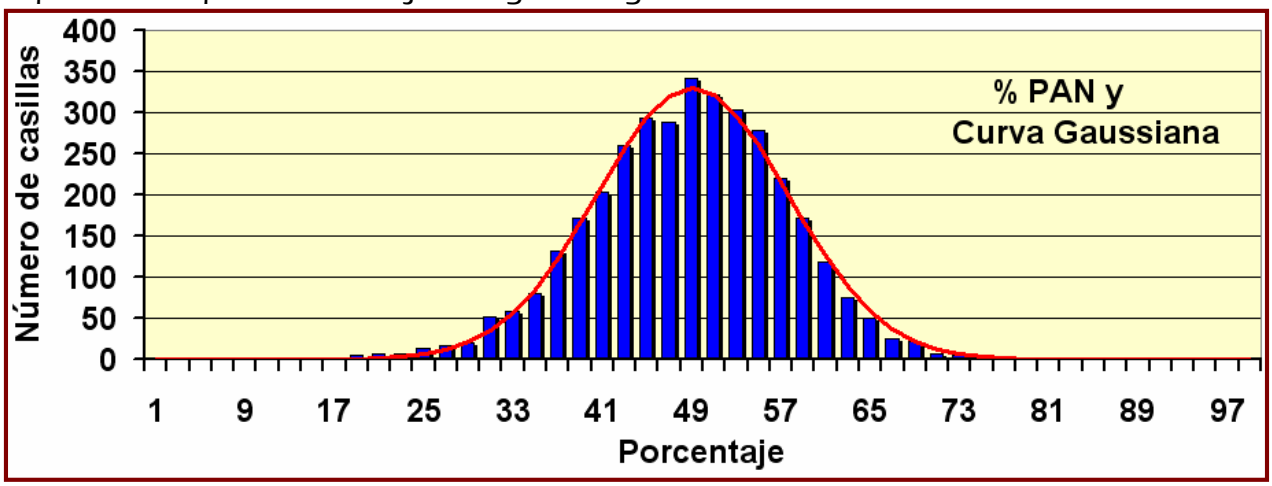
donde x es la variable independiente (los porcentajes en nuestro caso), y $f(x)$ es la altura de la función (número de casillas). Los parámetros μ y σ son –respectivamente- el valor promedio y la desviación estándar de la variable x. El significado de esta expresión será más claro con un ejemplo.

El ajuste a los datos de la votación

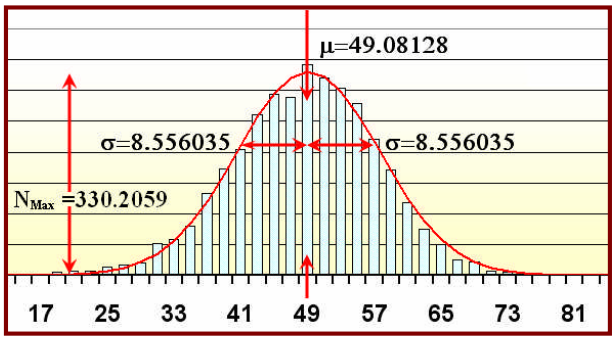
Si aplicamos la expresión anterior para representar los resultados de la elección en Baja California (*cf.* página 4.), usaríamos

$$N(p) = N_{Max} \times f(p) = \frac{N_{Max}}{\sigma\sqrt{2\pi}} e^{-(p-\mu)^2/2\sigma^2},$$

donde $N(p)$ nos dará el número de casillas que se espera que tengan un porcentaje p, y N_{Max} es una constante que se usa para establecer la altura de la curva. Usando $N_{Max} = 330.2059$, $\mu = 49.08128$ y $\sigma = 8.55603$, (sin ahondar por ahora en como fueron determinados estos valores), el número de casillas que esta expresión predice está representado por la curva roja en siguiente gráfica:



Obviamente, la curva gaussiana –ajustada a los datos de Baja California- nos da una muy buena representación de los resultados electorales.



Es fácil entender el significado de los valores usados viendo la gráfica de la izquierda: N_{Max} establece la altura de la curva, μ la posición del pico, y σ lo grueso de la campana.

Estos valores se pueden obtener aproximadamente usando el valor promedio del porcentaje, su desviación estándar y el número de

casillas correspondiente al porcentaje promedio. μ se puede aproximar por medio de un promedio *pesado:*

$$\mu = \frac{p_1 \times N_1 + p_3 \times N_3 + \cdots + p_{99} \times N_{99}}{N_1 + N_3 + \cdots + N_{99}}$$
$$= \frac{1\times 0 + \cdots + 47 \times 288 + 49 \times 342 + 51 \times 321 + \cdots + 99 \times 0}{0 + \cdots + 288 + 342 + 321 + \cdots + 0}$$
$$= \frac{172.553}{3549} = 48.62017$$

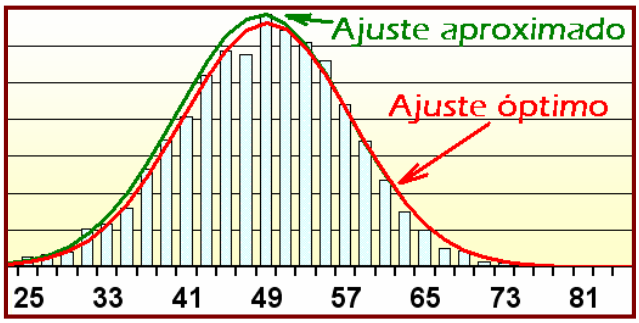
Y la desviación estándar por medio de

$$\sigma = \sqrt{\frac{(p_1 - \mu)^2 \times N_1 + \cdots + (p_{99} - \mu)^2 \times N_{99}}{N_1 + \cdots + N_{99}}}$$
$$= \sqrt{\frac{(1 - 48.62017)^2 \times 0 + \cdots + (49 - 48.62017)^2 \times 342 + \cdots + (99 - 48.62017)^2 \times 0}{0 + \cdots + 288 + 342 + 321 + \cdots + 0}}$$
$$= \sqrt{\frac{268312}{3549}} = \sqrt{75.60214} = 8.694949$$

Finalmente, N_{Max} es, aproximadamente, el número de casillas correspondiente a porcentaje $p = \mu = 48.62017 \rightarrow 49$, es decir es $N_{Max} \approx N(49) = 342$.

La función resultante con estos parámetros, sin embargo, no es la que se ajusta mejor a los datos de la elección, para obtener el ajuste óptimo, es necesario utilizar un procedimiento numérico de prueba y error que reduce la diferencia entre los datos y la gaussiana a un mínimo; este procedimiento es conocido como el método de mínimos cuadrados, y su explicación requiere de matemáticas más elevadas.

Usando tal método (el Lebenber-Marquart, para ser más específicos) se obtienen los valores mencionados inicialmente, es decir $N_{Max} = 330.2059$, $\mu = 49.08128$ y $\sigma = 8.55603$, y con estos se logra el ajuste representado por la curva roja. Una comparación de este ajuste con el que producen los valores aproximados ($N_{Max} \approx 342$, $\mu = 48.62017$, y $\sigma = 8.694949$) se puede ver en la gráfica de la derecha, donde la curva verde es la aproximada y la roja la del ajuste óptimo. Aunque por poco, la curva roja se ajusta mejor a los porcentajes de la votación.

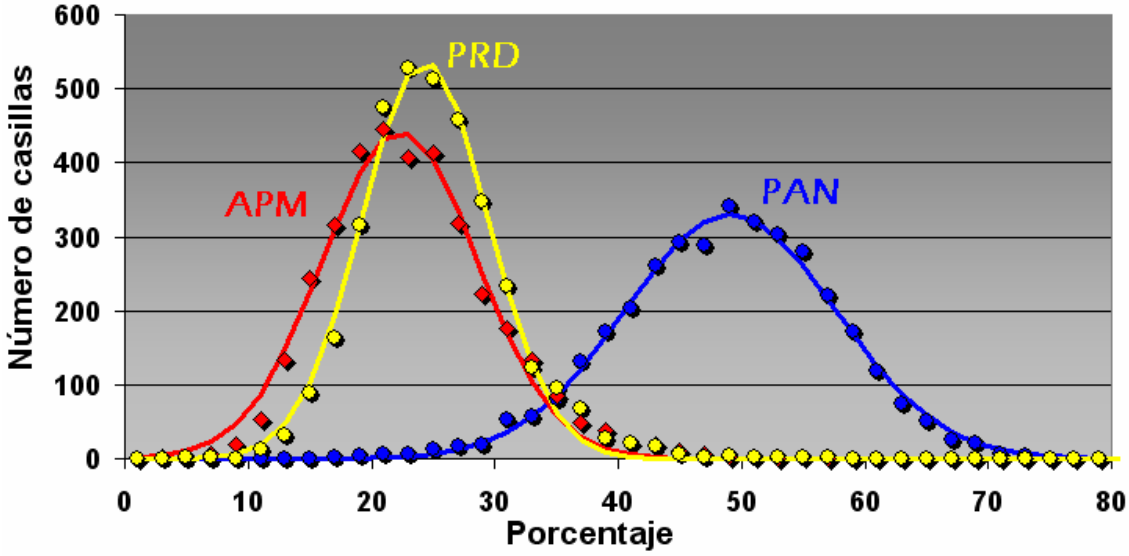


¿Y los otros partidos?

Repitiendo el procedimiento anterior con los votos recibidos por los demás partidos es posible obtener curvas gaussianas que representen esos resultados. Antes de hacerlo es útil recordar todos los pasos del proceso:

1. Transformación de número de votos en cada casilla a porcentajes.
2. Clasificación de porcentajes en clases de 2% para obtener la distribución de porcentajes.
3. Ajuste de la distribución resultante a una curva gaussiana por medio de la técnica numérica Lebenber-Marquart.

La siguiente figura muestra los resultados de los tres candidatos con mayor número de votos en el estado de Baja California. Para evitar congestionamiento de información, las barras de los porcentajes han sido reemplazadas por puntos.



La curva azul corresponde a la mostrada anteriormente (PAN), mientras que las nuevas son las distribuciones de votos obtenidas por la Alianza por México (APM) y por la Coalición por el Bien de Todos (PRD). Nótese una vez más que el "área" bajo las curvas corresponden al número total de casillas, como se puede verificar con la aproximación del triángulo usada anteriormente.

Resultados inmediatos son que

i) A pesar de que el PRI gobierna Tijuana, quedó en último lugar en el estado.
ii) A pesar de no gobernar ninguna ciudad grande en B.C., el PRD se llevó una cuarta parte de los votos,
iii) El PAN arrasó con la mitad de los votos.
iv) Las tres curvas son descritas fielmente por gaussianas.

Este último punto da la pauta para continuar con este análisis. La votación estatal de un partido puede ser descrita por tres números: N_{Max}, μ, y σ; desviaciones pronunciadas de este comportamiento gaussiano representarían anomalías que tendrían que ser explicadas.

Ahora repitamos el estudio para todos los estados.

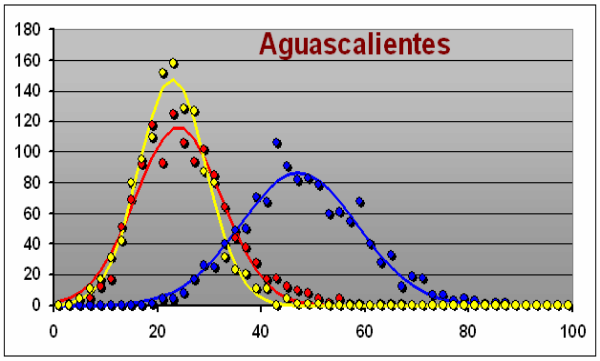
¿Y los demás estados?

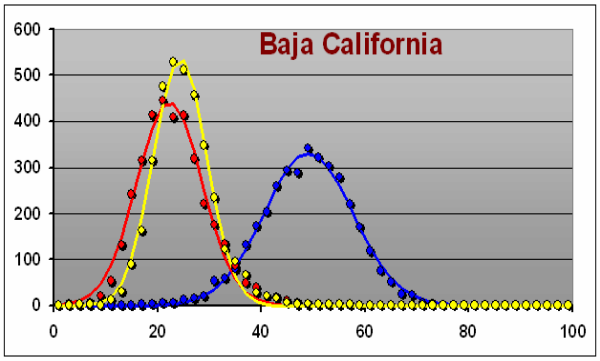
Aunque los resultados presentados en la sección anterior son característicos de los demás estados, existen diferencias importantes. Las gráficas de las siguientes páginas muestran las distribuciones de los porcentajes de votos para las 32 entidades federativas del país. El código de colores usado es: azul para el PAN, rojo para el PRI y amarillo para el PRD.

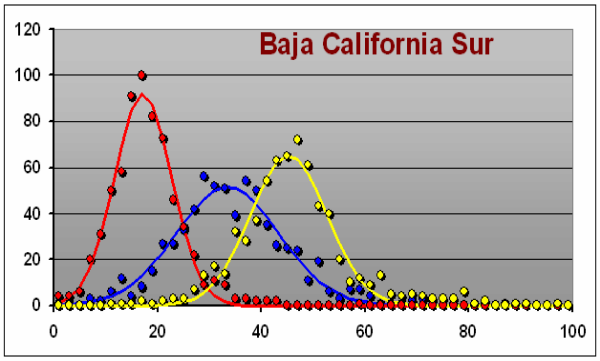
Estados que presentan irregularidades –tales como curvas incompletas, desviaciones obvias del comportamiento gaussiano, etc.— son

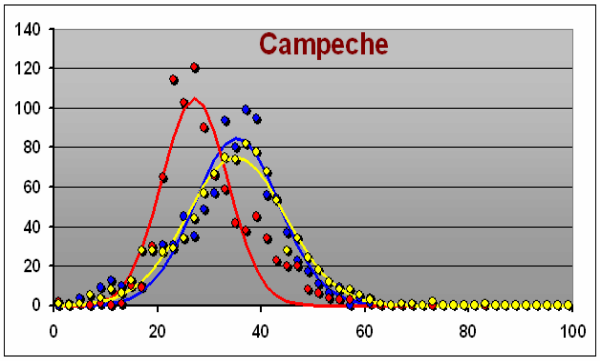
- PRI en Campeche, Chiapas, Hidalgo, Oaxaca, Puebla y San Luis Potosí
- PAN en Chiapas y Oaxaca
- PRD en Sonora

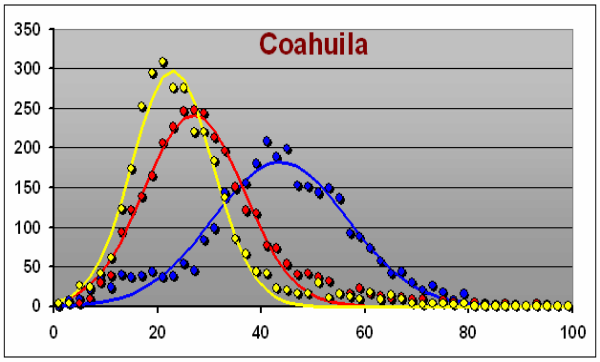
Estas diferencias pueden ser cuantificadas por medio del método siguiente.

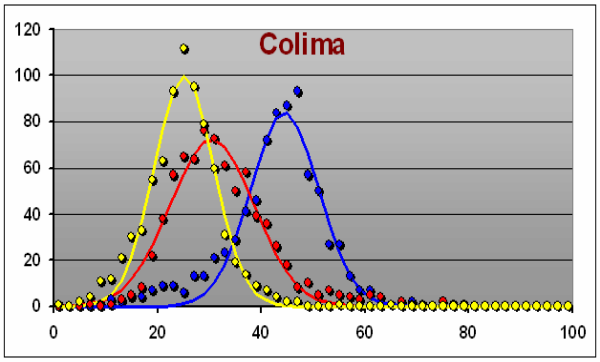

Aguascalientes
180
160
140
120
100
80
60
40
20
0
0
20
40
60
80
100

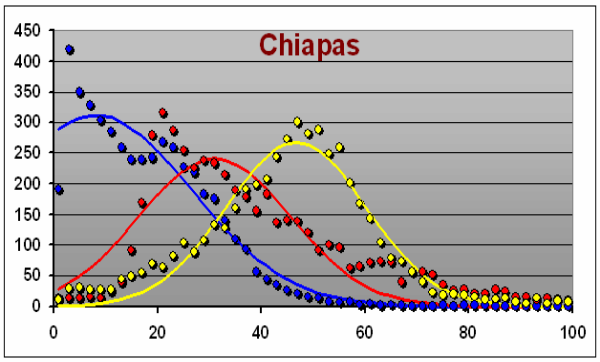

Baja California
600
500
400
300
200
100
0
0
20
40
60
80
100

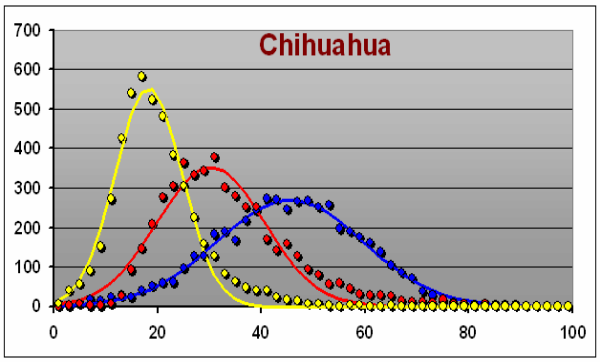

Baja California Sur
120
100
80
60
40
20
0
0
20
40
60
80
100

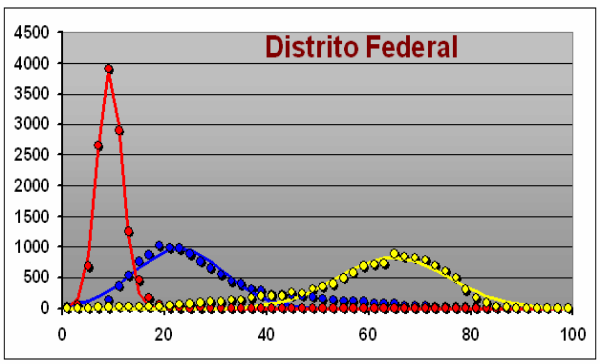

Campeche
140
120
100
80
60
40
20
0
0
20
40
60
80
100

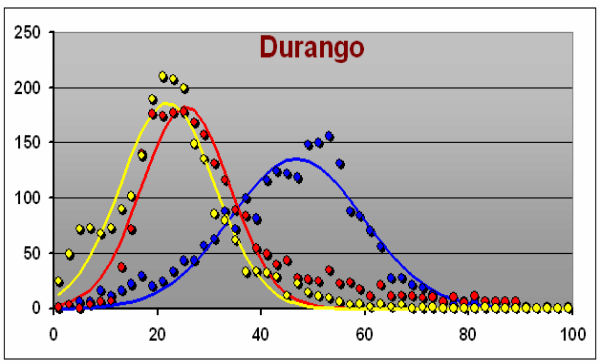

Coahuila
350
300
250
200
150
100
50
0
0
20
40
60
80
100

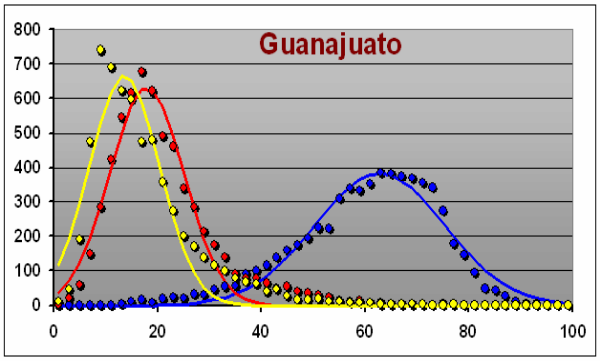

Colima
120
100
80
60
40
20
0
0
20
40
60
80
100

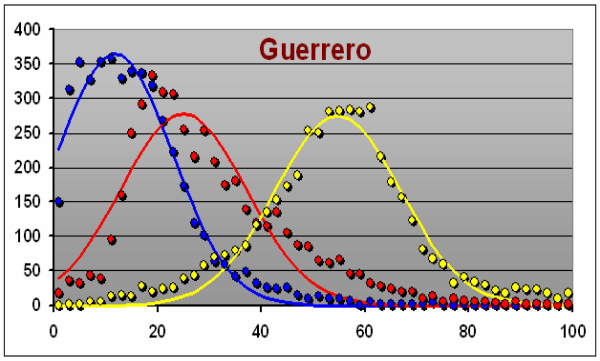

Chiapas
450
400
350
300
250
200
150
100
50
0
0
20
40
60
80
100

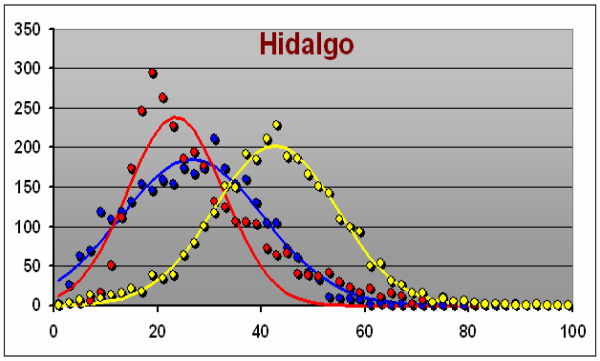

Chihuahua
700
600
500
400
300
200
100
0
0
20
40
60
80
100

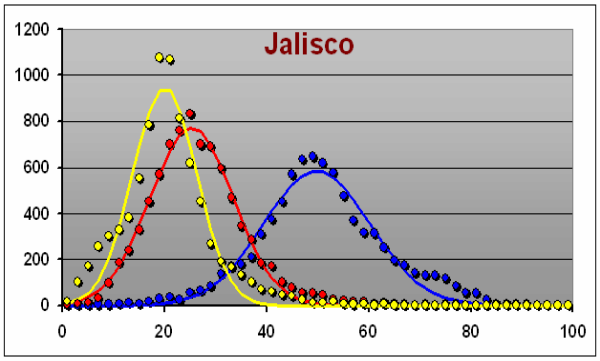

Distrito Federal
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
20
40
60
80
100

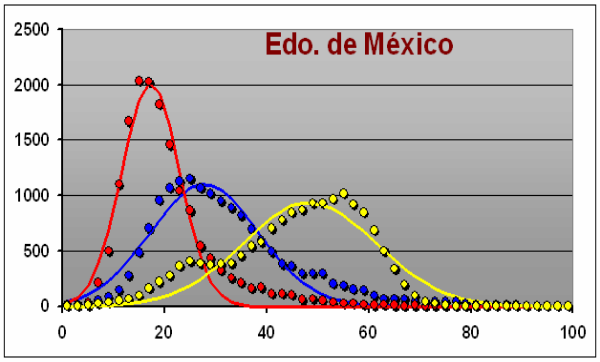

Durango
250
200
150
100
50
0
0
20
40
60
80
100

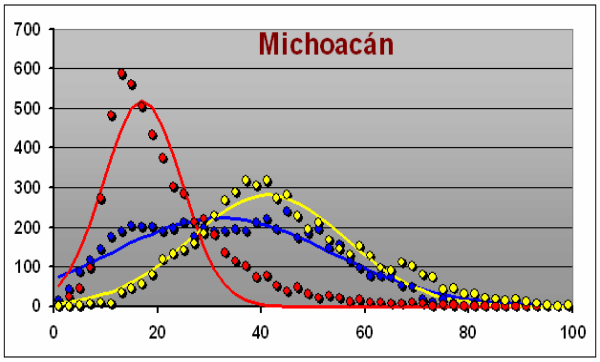

Guanajuato
800
700
600
500
400
300
200
100
0
0
20
40
60
80
100

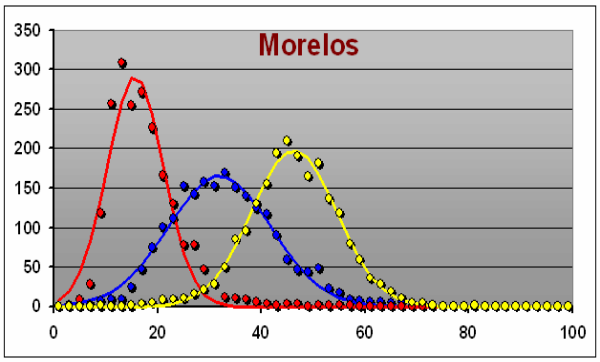

Guerrero
400
350
300
250
200
150
100
50
0
0
20
40
60
80
100

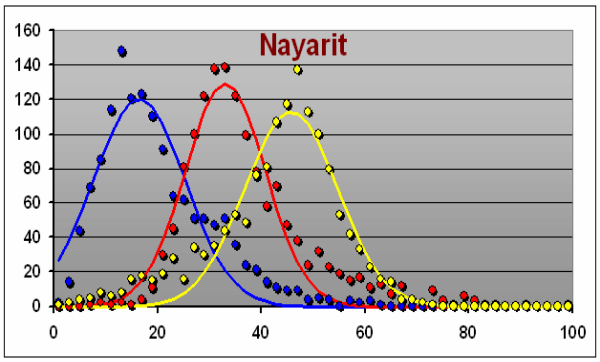

Hidalgo
350
300
250
200
150
100
50
0
0
20
40
60
80
100

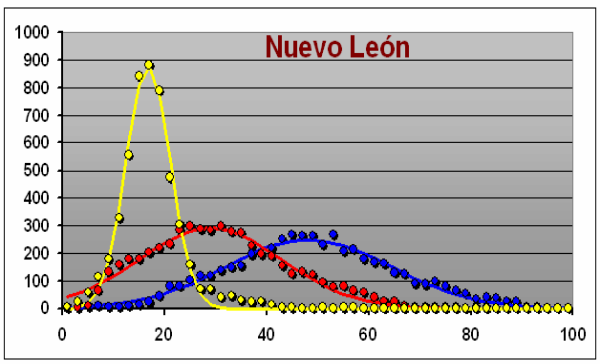

Jalisco
1200
1000
800
600
400
200
0
0
20
40
60
80
100

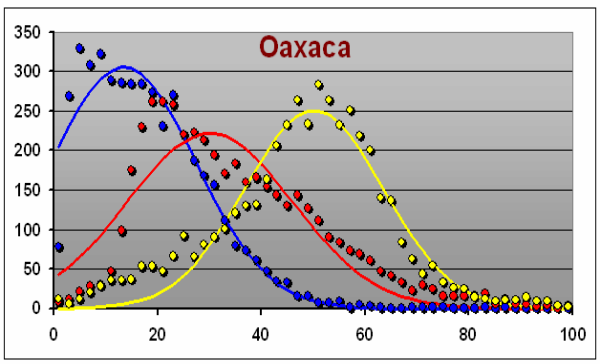

Edo. de México
2500
2000
1500
1000
500
0
0
20
40
60
80
100

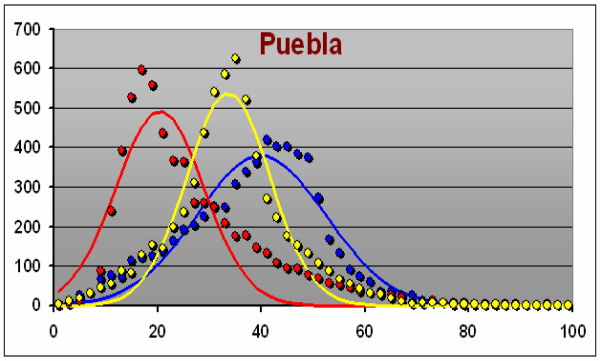

Michoacán
700
600
500
400
300
200
100
0
0
20
40
60
80
100

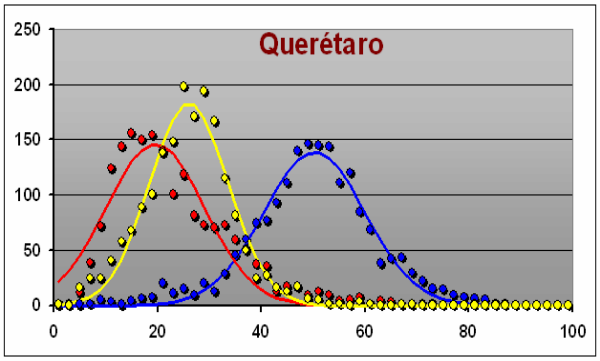

Morelos
350
300
250
200
150
100
50
0
0
20
40
60
80
100

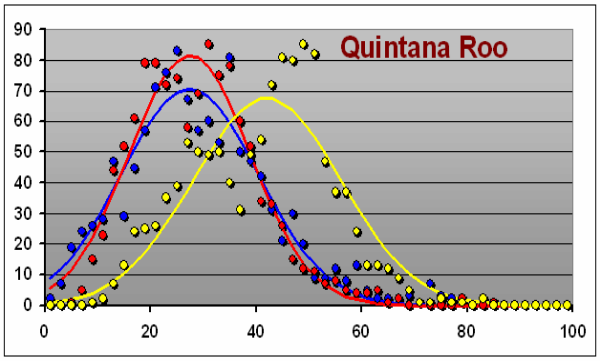

Nayarit
160
140
120
100
80
60
40
20
0
0
20
40
60
80
100

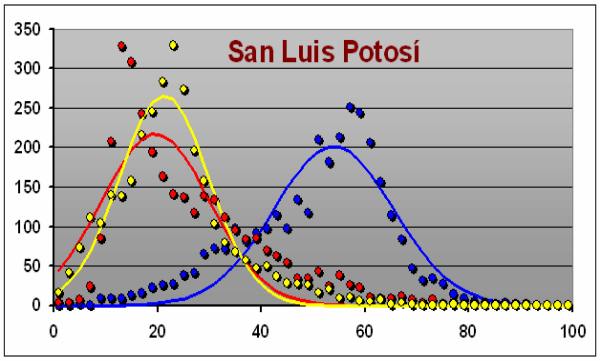

Nuevo León
1000
900
800
700
600
500
400
300
200
100
0
0
20
40
60
80
100

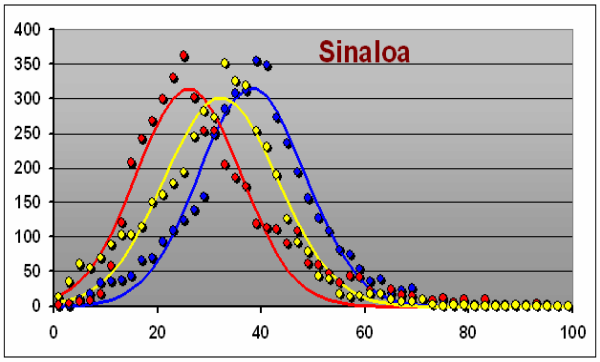

Oaxaca
350
300
250
200
150
100
50
0
0
20
40
60
80
100

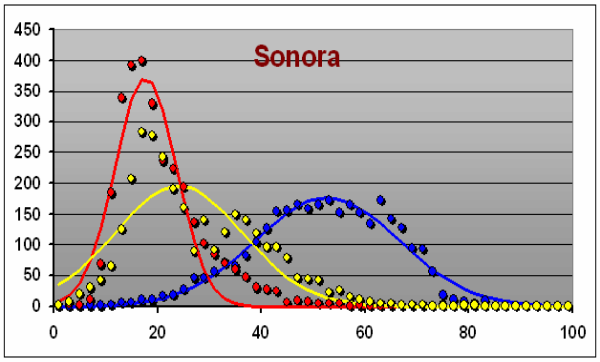

Puebla
700
600
500
400
300
200
100
0
0
20
40
60
80
100

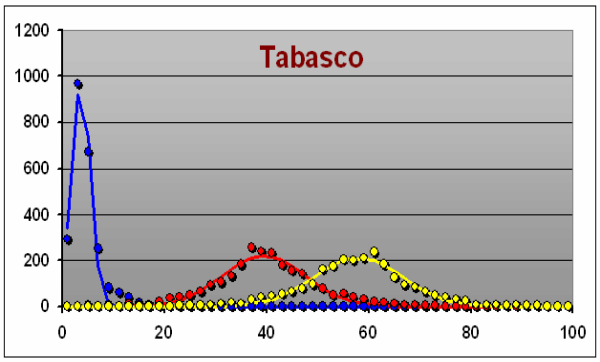

Querétaro
250
200
150
100
50
0
0
20
40
60
80
100

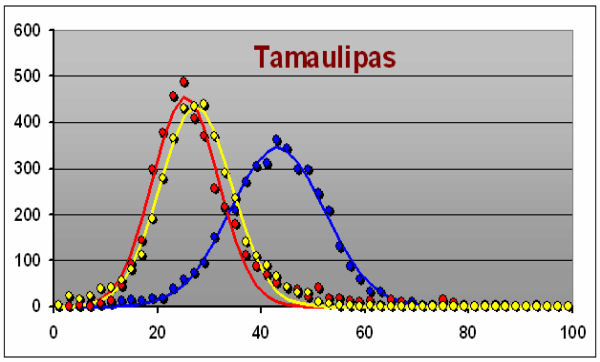

Quintana Roo
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100

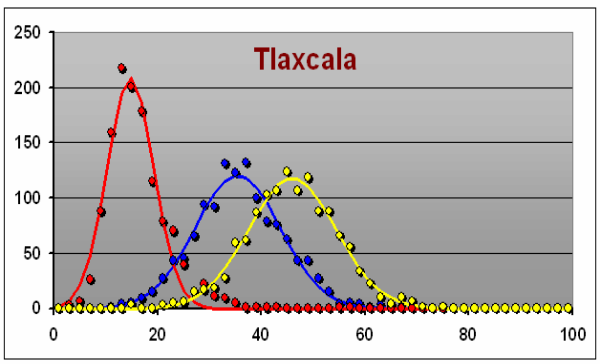

San Luis Potosí
350
300
250
200
150
100
50
0
0
20
40
60
80
100

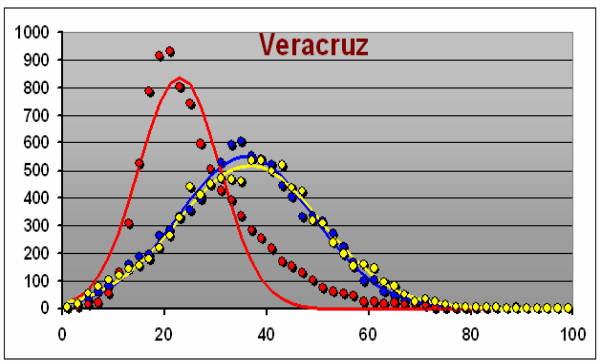

Sinaloa
400
350
300
250
200
150
100
50
0
0
20
40
60
80
100

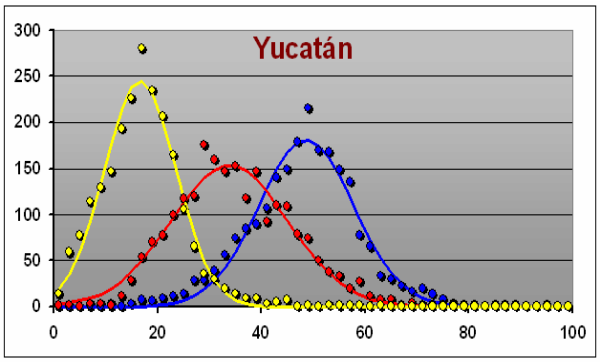

Sonora
450
400
350
300
250
200
150
100
50
0
0
20
40
60
80
100

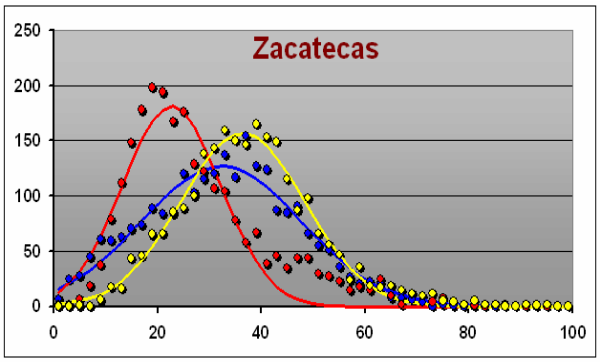

Tabasco
1200
1000
800
600
400
200
0
0
20
40
60
80
100


Tamaulipas
600
500
400
300
200
100
0
0
20
40
60
80
100


Tlaxcala
250
200
150
100
50
0
0
20
40
60
80
100

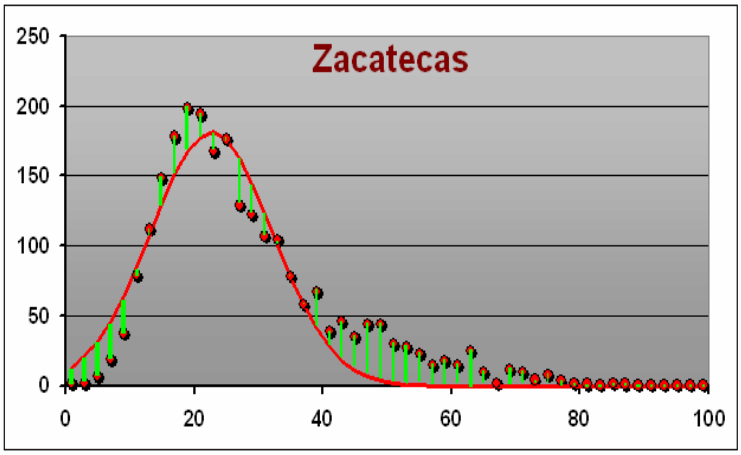

Veracruz
1000
900
800
700
600
500
400
300
200
100
0
0
20
40
60
80
100





Midiendo la falta de gaussianidad

Es posible medir la desviación de los datos respecto a la curva gaussiana. Tomando como ejemplo la distribución de los votos del PRI en Zacatecas, podemos ilustrar las diferencias entre los datos y la curva usando las líneas verdes mostradas; estas representan las diferencias en número de casillas. La suma de todas estas diferencias servirá como medida de la desviación.

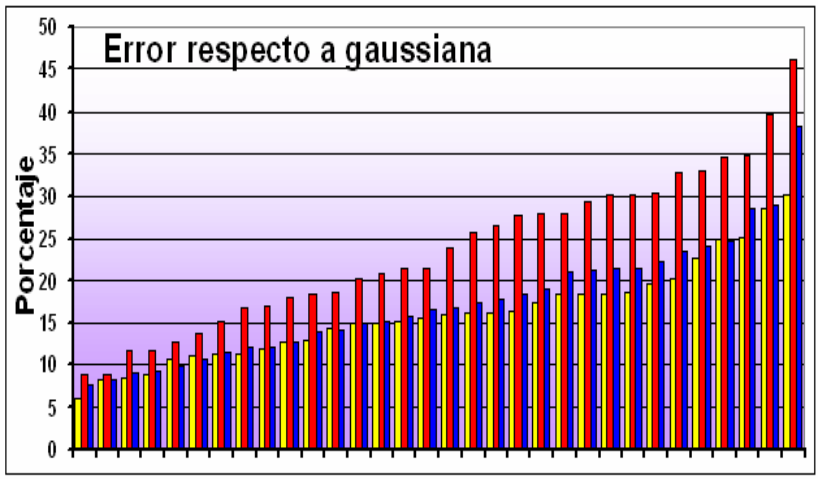


Debido a que diferentes estados tienen diferente números de casillas, es preferible expresar esa diferencia en porcentaje del total de casillas para cada estado. Asimismo, para evitar que un error negativo cancele a uno positivo, conviene sumar los valores absolutos de estas diferencias.

Matemáticamente, si representamos el número de casillas que obtuvieron un porcentaje p_i por $N^D(p_i)$, y el número correspondiente estimado por la distribución gaussiana como $N(p_i)$, la diferencia entre los datos y la curva puede ser cuantificada por medio de

$$D = \frac{\left|N^D(p_1) - N(p_1)\right| + \cdots + \left|N^D(p_{99}) - N(p_{99})\right|}{N^D(p_1) + \cdots + N^D(p_{99})} \times 100$$

En palabras, D mide el error entre los datos y curva de gauss como un porcentaje del número total de casillas; la ventaja de usar porcentajes es la de poder combinar resultados de todos los estados.



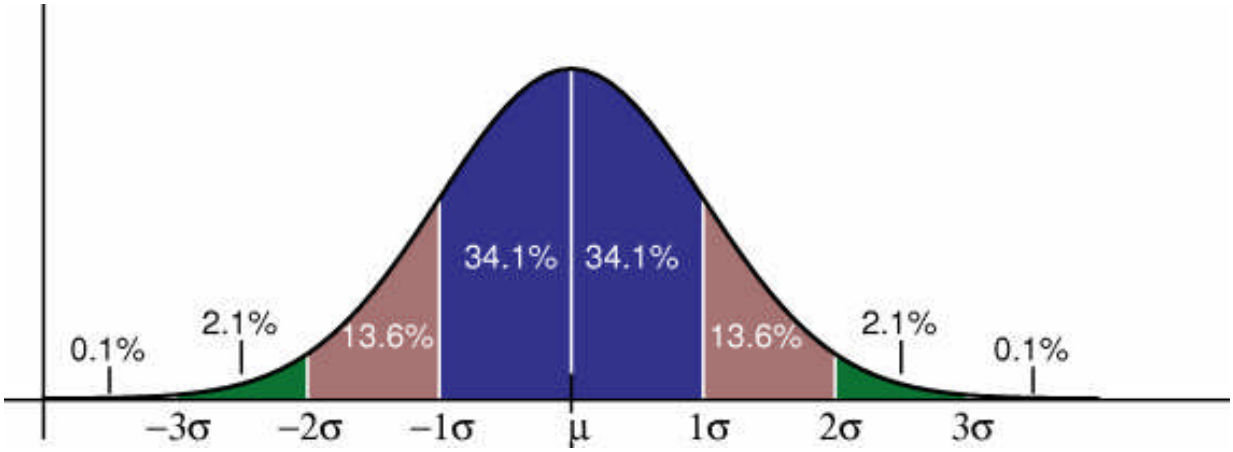
Aplicando esta medida a los estados nos da los resultados de la gráfica de la derecha que muestra los errores de los

tres partidos en los 32 estados y Distrito Federal ordenados, no por estado, sino de mayor a menor para facilitar una comparación entre partidos; de nuevo rojo representa PRI, azul PAN y amarillo PRD. La observación inmediata –de que el PRI difiere más que los demás de las predicciones gaussianas- es ratificado por el promedio de tales errores: 16.1 % para el PAN, 17.4 % para el PRD, y 23.6 % para el PRI. Nuestro siguiente paso en este análisis será determinar que tan probables son estas diferencias.

Midiendo la probabilidad de los errores

La ventaja de trabajar con distribuciones gaussianas es que muchas de sus propiedades son conocidas. *Normalidad*, es decir el hecho de que los datos se ajusten a una gaussiana, se ha convertido en la suposición central de la teoría de errores. Esta teoría, a su vez demuestra que las diferencias entre los datos y un ajuste normal deben tener una distribución normal; esta propiedad permite evaluar la probabilidad de ocurrencia de estas diferencias.

En otras palabras, cuando se sabe que una distribución sigue un comportamiento gaussiano, se debe esperar que las diferencias entre la distribución y un ajuste gaussiano estén -a su vez– distribuidas de manera gaussiana.



Así pues, estas diferencias deberán tener una distribución como la mostrada en la figura (la cual es para una distribución normal, es decir gaussiana con $\mu = 0$ y $\sigma = 1$). El 68.2 % de estas diferencias deberán ser chicas, alrededor de $\mu = 0$ en la región delimitada por $\mu - \sigma$ y $\mu + \sigma$ (zona azul). Otras diferencias tendrán las siguientes probabilidades de existir:

- 13.6 % tendrán magnitudes entre $\mu + \sigma$ y $\mu + 2\sigma$, o entre $\mu - 2\sigma$ y $\mu - \sigma$ (zona café).
- 2.1 % estarán entre $\mu + 2\sigma$ y $\mu + 3\sigma$, o entre $\mu - 3\sigma$ y $\mu - 2\sigma$ (zona verde).
- 0.1 % entre $\mu + 3\sigma$ y $\mu + 4\sigma$, o entre $\mu - 4\sigma$ y $\mu - 3\sigma$.

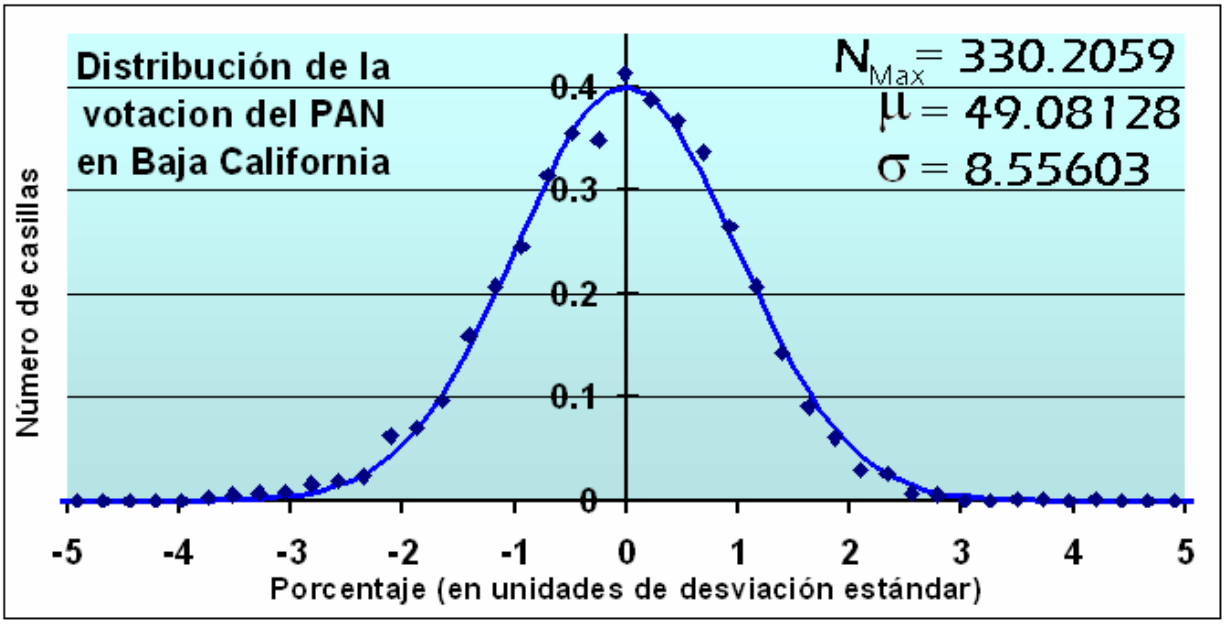
Otra manera de expresar esto es:

- 95 de cada 100 diferencias tendrán magnitudes entre $\mu - 2\sigma$ y $\mu + 2\sigma$.
- 1 de cada 100 diferencias tendrá una magnitud mayor que $\mu + 2.3\sigma$.
- 1 de cada 1,000 diferencias tendrá una magnitud mayor que $\mu + 3.05\sigma$.
- 1 de cada 10,000 diferencias tendrá una magnitud mayor que $\mu + 3.7\sigma$.
- 1 de cada 100,000 diferencias tendrá una magnitud mayor que $\mu + 4.2\sigma$.
- 1 de cada 1,000,000 diferencias tendrá una magnitud mayor que $\mu + 4.75\sigma$.
- 1 de cada 10 millones diferencias tendrá una magnitud mayor que $\mu + 5.4\sigma$.

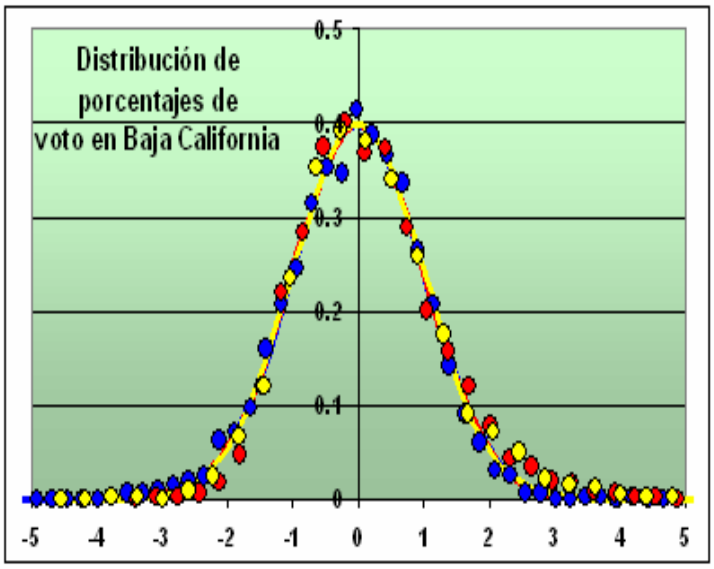
Obviamente, como tan solo tenemos 50 valores de porcentajes por estado, y un total de 50x32 = 1600 valores en todo el país, es de esperarse que no veamos diferencias mayores de, digamos, $\mu + 3.2\sigma$.

Normalizando los porcentajes

Para estudiar la distribución de la diferencias en todo el país es necesario combinar resultados de todos los estados, esto requiere que los datos de cada estado estén *normalizados*, es decir que se ajusten a una gaussiana con $\mu = 0$ y $\sigma = 1$. Esta normalización se logra transformando los porcentajes p_i en $Z_i = (p_i - \mu)/\sigma$, donde μ y σ son la media y desviación estándar de la distribución gaussiana a transformar. Si a esta transformación le añadimos una reducción de escala dividiendo por N_{Max}, las distribuciones resultantes serán gaussianas con $\mu = 0$ y $\sigma = 1$, y altura máxima de $1/\sqrt{2\pi\sigma^2}$, es decir una *normal estándar*, las cuales podrán ser combinadas entre si. Un ejemplo de este procedimiento se muestra en la figura siguiente.



Transformando los valores de Baja California (ver gráfica en página 6), con $N_{Max} = 330.2059$, $\mu = 49.08128$ y $\sigma = 8.55603$, se obtiene la curva normal estándar mostrada.

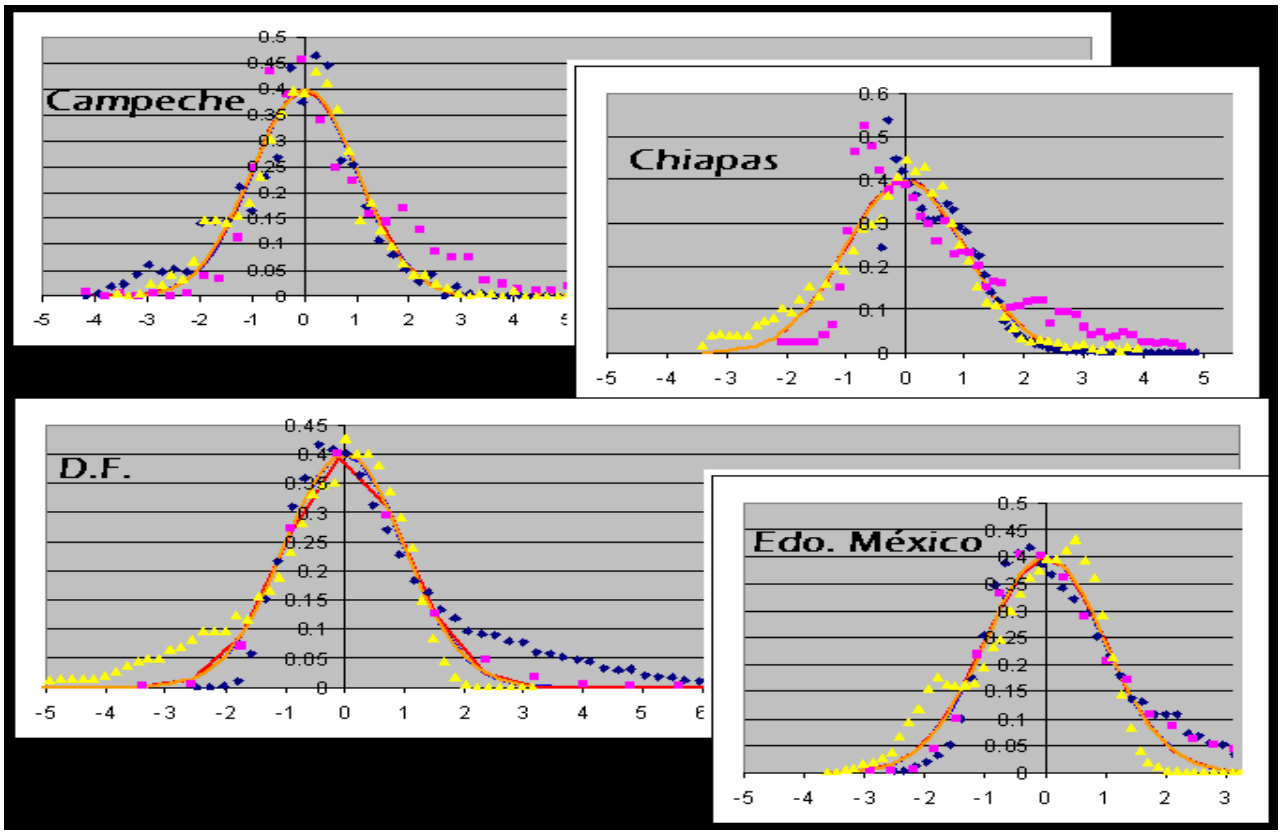


Repitiendo para los demás partidos se obtiene la siguiente gráfica que corresponde a la de la página 8. Hay que notar que la gráfica contiene tres curvas sobrepuestas, una para cada partido.

Repitiendo este procedimiento para todos los estados se obtienen curvas similares con las que se

puede hacer el estudio de las diferencias. Revisando algunos casos, como los mostrados en la siguiente gráfica, es posible ver que el uso de curvas normalizadas pone en manifiesto la existencia de errores que antes –con las gráficas sin normalizar- eran imperceptibles.

Por ejemplo, comparando la curva normalizada del D.F. con la original (ver página 9) se puede detectar la existencia de unas colas sumamente pronunciadas para el PRD (exceso en porcentajes bajos) y el PAN (exceso en porcentajes altos).



Estudio de errores

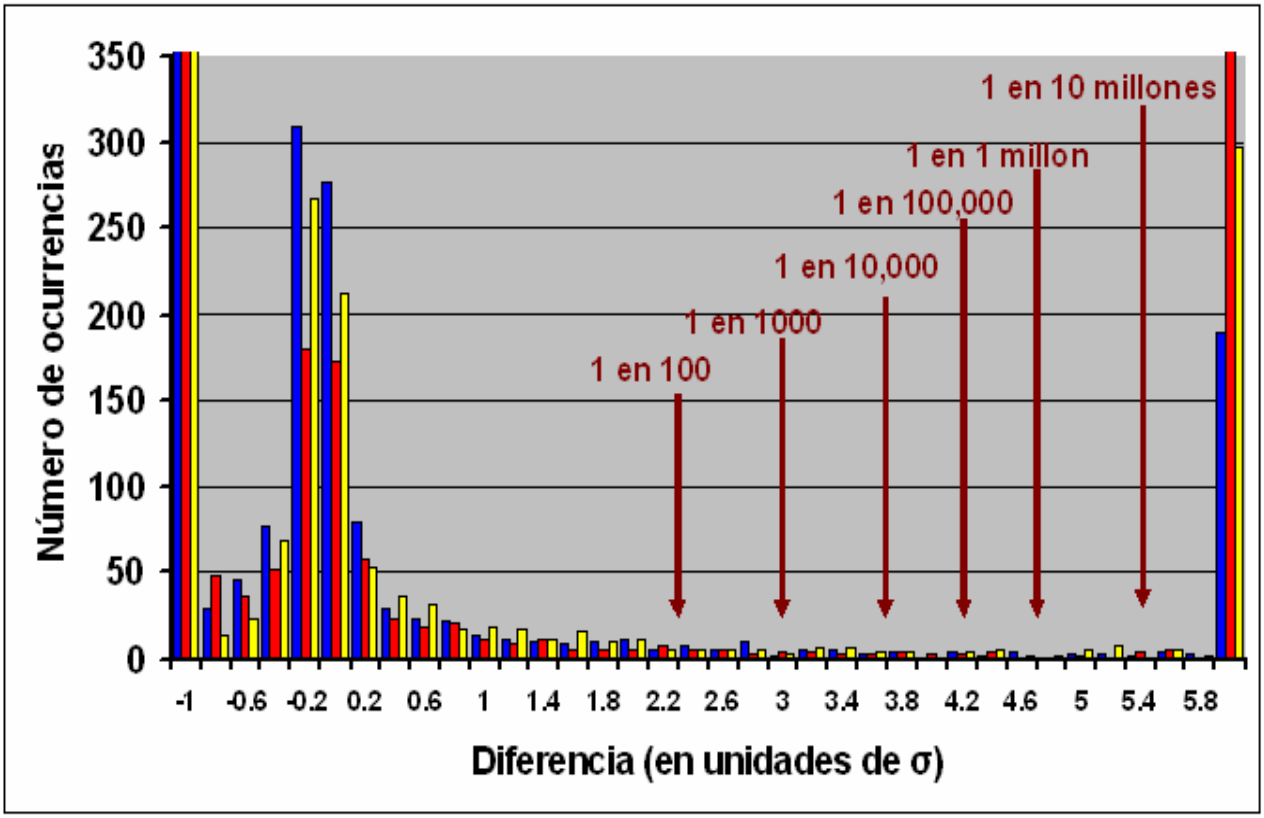
Para entender el estudio de errores es útil –de nuevo- recordar todos los pasos del proceso que nos ha traído hasta este punto:

1. Los votos recibidos por cada partido en cada casilla fueron transformados en porcentajes.
2. Estos porcentajes fueron clasificados en clases de 2% para obtener la distribución de porcentajes.
3. La distribución de porcentajes resultante fue ajustada a una curva gaussiana por medio de la técnica numérica Lebenber-Marquart.
4. Las distribuciones y los ajustes gaussianos fueron transformados en distribuciones normales estándar con $\mu = 0$ y $\sigma = 1$.

Es sobre estas curvas normales que procedemos a contar las diferencias entre los datos y los ajustes. A diferencia con el procedimiento seguido en la página 12, para que las diferencias den a su vez una distribución normal estándar (es decir con $\mu = 0$ y $\sigma = 1$), es necesario expresar las diferencias en unidades del valor esperado, es decir cada diferencias se calculará por medio de

$$d_i = \frac{\mathcal{N}^D(p_i) - \mathcal{N}(pi)}{\mathcal{N}(p_i)}$$

donde $\mathcal{N}^D(p_i)$ es el valor normalizado del número de casillas con porcentaje p_i, y $\mathcal{N}(p_i)$ es el valor normalizado del ajuste gaussiano; de esta manera, las diferencias estarán dadas en unidades de desviaciones estándar, $\sigma = 1$. Clasificando estas diferencias en clases de 0.2σ, se puede obtener la distribución de las mismas y así asignarles probabilidades de ocurrir.



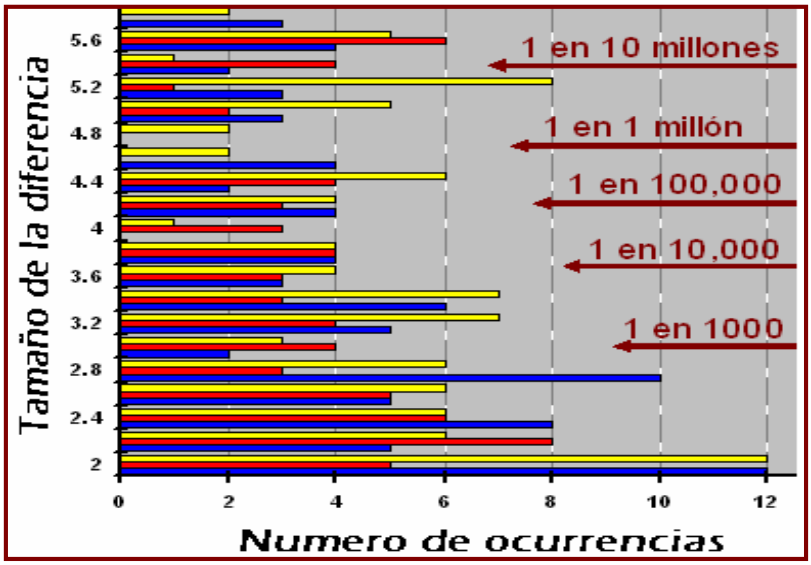
La gráfica superior muestra la distribución obtenida para los 1600 datos de las 32 entidades federativas. Sin tomar en cuenta los picos iniciales y finales, el resto sí corresponde a la curva normal esperada. Varios detalles requieren explicación.

El pico inferior. Debido a que el número de casillas es siempre positivo o cero, tenemos que $\mathcal{N}^D(p_i) \geq 0$, por lo que d_i calculada con $d_i = \left(\mathcal{N}^D(p_i) - \mathcal{N}(pi)\right) / \mathcal{N}(p_i)$ tendrá como valor mínimo de -1, es decir $d_i \geq -1$; esto introduce una terminación artificial de la curva en $d_i = -\sigma$ y un acumulamiento en la primera clase correspondiente a diferencias con valores entre $-\sigma \leq d \leq -0.8\sigma$. Estos picos, que exceden el máximo de la escala de la gráfica, tienen valores de 376, 381 y 408 para el PAN, PRI y PRD, correspondientemente.

El pico superior. Muy en contra de lo esperado, existen diferencias mayores de 5 y 6 desviaciones estándar. Para no extender el límite derecho de la gráfica, se acumularon todos los valores mayores de $d_i \geq 6\sigma$ en una sola clase resultando en el pico mostrado. De nuevo, con valores de 190 (PAN), 497 (PRI) y 298 (PRD), estos picos exceden el máximo de la escala.

Probabilidad de errores

Como fue explicado en la página 13, graficando los datos de esta manera nos permite estimar la probabilidad de que ocurra algún valor de las diferencias.



La gráfica de la izquierda muestra una ampliación rotada de la anterior. Esta ampliación excluye diferencias menores a 2σ y mayores a 5.8σ.

Las flechas –igual que en la gráfica anterior- indican las probabilidades de obtener diferencias mayores que esos límites. Por ejemplo, la probabilidad de que se encuentren diferencias mayores a 3.05σ es de 1 en mil, o 0.001; para nuestro caso de 1600 datos, arriba de esta flecha debería de haber tan sólo 16 casos, pero existen 263 (PAN), 560 (PRI) y 393 (PRD).

Diferencia	Probabilidad	Numero esperado	Encontrados		
			PAN	PRI	PRD
> 2.2 σ	1/100	160	263	560	393
> 3.05 σ	1/1000	1.6	233	534	356
> 3.7 σ	1/10,000	0.16	219	524	338
> 4.2 σ	1/100,000	0.016	211	514	329
> 4.75 σ	1/1,000,000	0.0016	205	510	321
> 5.4 σ	1/10,000,000	0.00016	199	507	306

En resumen, encontramos cientos de casos donde las diferencias entre los datos y lo esperado para una distribución normal son sumamente grandes. En porcentajes, -y aceptando los resultados hasta ≈5 σ como válidos, se podría afirmar que 12.4 % de los porcentajes del PAN (12.4=199/1600x100), 31.68 % del PRI, y 19 % del PRD son totalmente anómalos y no corresponden a lo esperado.

Conclusión

Se analizaron los datos de la elección presidencial. Se observó que los porcentajes de datos obtenidos siguen una distribución gaussiana estado por estado, lo cual permitió usar estadística conocida para medir la *normalidad* de la elección e identificar desviaciones de este patrón. En el análisis de errores se detectaron desviaciones significativas que ponen en duda la pureza de los datos. Resultados como los obtenidos (por ejemplo, para diferencias mayores a 5.4 σ) se verían una vez en cada 6250 elecciones (=10,000,000/1600), es decir una vez cada 37,500 años.