INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

Blah blah blah blah blah blah

T E S I S

QUE PARA OBTENER EL TÍTULO DE

MAESTRA EN CIENCIA DE DATOS

P R E S E N T A

MARIANA CARMONA BAEZ

ASESOR DE TESIS:
M.E. María Teresa Ortiz Mancera

México, D.F. 2020

Con fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada **"Blah blah blah blah blah blah"**, otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Bailleres Jr. autorización para que fijen la obra en cualquier medio, incluido el electrónico y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por la divulgación una contraprestación.

MARIANA CARMONA BAEZ

_____

FECHA

_____

FIRMA

Para mi Aldito.

Agradecimientos

A mis padres.

Índice general

Capítulo 1

Introducción

Capítulo 2

Inferencia Bayesiana

Un poco de historia.

2.1. Flujo de trabajo Bayesiano

La idea de tener un flujo de trabajo Bayesiano es poder desarrollar modelos robustos para resolver problemas prácticos.

2.1.1. Paso 1: Análisis conceptual

2.1.2. Paso 2: Definir las observaciones

2.1.3. Paso 3: Identificar estadísticas resumen relevantes

2.1.4. Paso 4: Construir un modelo

2.1.5. Paso 5: Identifcar nuevas estadísticas resumen

2.1.6. Paso 6: Analizar el ensamble

2.1.7. Paso 7: Ajustar las observaciones

2.1.8. Paso 8: Analizar la distribución predictiva posterior

Capítulo 3

Flujo de trabajo Bayesiano

Tomando en cuenta lo anterior, se puede diseñar un flujo de trabajo bayesiano que guíe el desarrollo de modelos robustos en aplicaciones prácticas. El flujo de trabajo comienza examinando el diseño experimental y el proceso de medición resultante. Solo una vez que se comprende el proceso de medición, construimos nuestro modelo inicial y estudiamos su desempeño en el contexto de nuestra experiencia estadística y de dominio. Finalmente, podemos analizar el ajuste del modelo en los datos observados, verificando la precisión computacional del ajuste y la adecuación del modelo para capturar las características relevantes del verdadero proceso de generación de datos.

Aquí, la experiencia en el dominio se refiere a la experiencia de los responsables de recopilar, curar o manipular los datos, así como de las partes interesadas que tomarán decisiones utilizando las inferencias finales o cualquier intermediario que ayude a tomar esas decisiones. La experiencia estadística se refiere a la competencia en modelado probabilístico y computación.

En este trabajo se eligió el tema de la participación electoral en las elecciones presidenciales en México para ejemplificar el flujo de trabajo Bayesiano.

3.1. Diseño experimental

Antes de empezar a modelar se tiene que considerar el diseño del experimento en el contexto en el que se está trabajando.

3.1.1. Análisis conceptual

El análisis conceptual sirve para obtener información de cómo fueron generadas las observaciones. Se necesita comprender el proceso de medición para poder identificar efectos sistemáticos que influyen en dicho proceso pero que no se conocen con certeza. Este análisis forma la base del modelo.

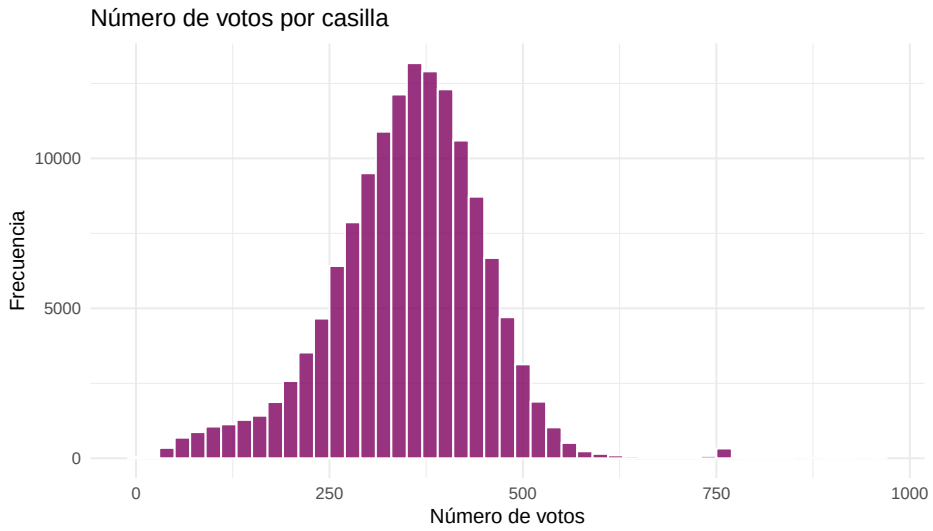
Ya que el objetivo de este ejemplo es modelar la participación ciudadana en las elecciones presidenciales de 2018 en México utilizando los datos del conteo rápido, se empezará por describir el contexto que rodea la participación electoral, así como la descripción de lo que es el conteo rápido.

3.1.1.1. Participación Electoral

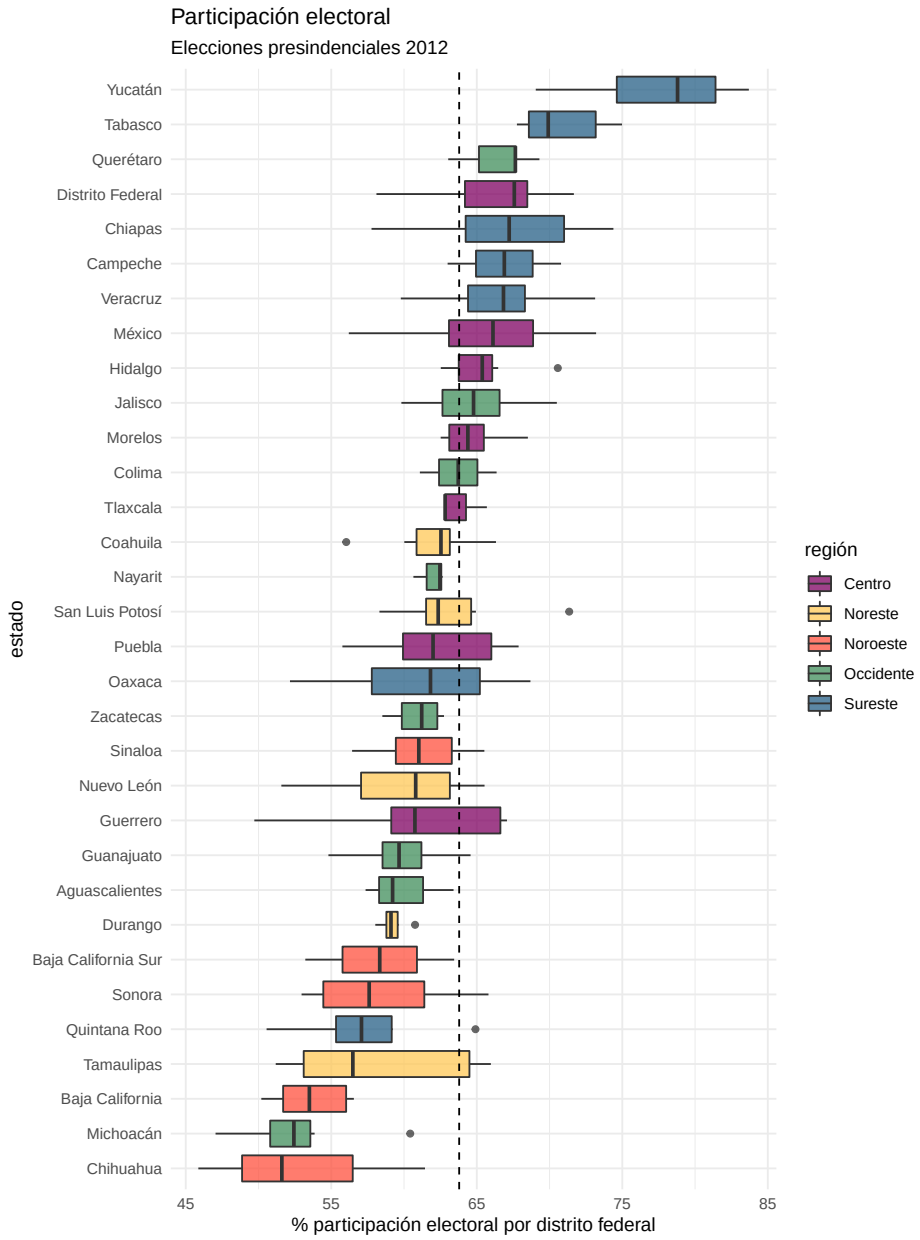
La participación de los ciudadanos en las elecciones es una parte crucial de una democracia, democracia multiparidista en el caso de México. Actualmente no existe un consenso acerca de los niveles de participación ciudadana necesarios para mantener una democracia.

Alrededor del mundo se ha visto que varias democracias han presentado disminusiones en las tasas de participación electoral. Este fenómeno puede deberse a varios factores como: la falta de confianza en las instituciones, procesos de globalización, o bien la necesidad de nuevos modelos de participación. En el caso particular de América Latina, a la par del crecimiento del índice de desarrollo humano y la consolidación de sistemas democráticos, se ha presentado un crecimiento en la participación electoral. Sin embargo, los niveles de participación de México se encuentran por debajo del promedio de la región. En 2006, la tasa de participación en México fue de 55 %, mientras que en América Latina fue de 66 %.

Para las elecciones presidenciales del 2012 en México, se vio una alza en la participación electoral de las elecciones presidenciales (63.34 %). No obstante, este porcentaje sigue siendo inferior a los niveles registrados en 1994 (77 %). El pocentaje de participación para 2018 se mantuvo en el mismo nivel que en 2012 (63 %).



Para entender el contexto en el cual ocurrieron las elecciones de 2018, se toma como referencia los datos de las elecciones presidenciales de 2012. Se observa que los niveles de participación electoral más altos corresponden a los estados de Yucatán y Tabasco con 77 y 71 por ciento de ciudadanos que acudieron a las urnas, respectivamente. Los niveles más bajos pertenecen a los estados de Michoacán, Chihuahua, y Baja California con un nivel de votación en torno a 53 %. Los demás estados presentan un nivel de participación entre 58 y 67 por ciento. En la siguiente gráfica también puede hablarse de un patrón por región geográfica. Los estados del Sureste y Centro del país tienen un nivel de participación mayor a la mediana nacional en la mayoría de los casos.



En la elección de 2018, los estados de Yucatán y Tabasco permanecen en la cima de este indicador, el primero con un nivel de participación de 75 % y el segundo de 71 %, es decir, con una diferencia de -2 y 0 puntos porcentuales, respectivamente, en comparación con la elección anterior. No obstante, en esta ocasión se acercan a estas proporciones la Ciudad de México y el estado de Campeche, en cada uno de estos casos, siete de cada 10 ciudadanos acudieron a votar.

En este mismo año (2018) Baja California y Chihuahua continúan como dos de los estados con menor participación en el país, pero esta vez se les suman Sonora y Guanajuato. Los cuatro en torno a 53 %, es decir, una proporción similar a la de 2012. Los demás estados se

encuentran en un rango que va de 56 a 68 por ciento. En México, generalmente seis de cada 10 mexicanos emiten su voto.

(mapa de la participación)

Ahora que se está en el contexto, se definirá lo que es el conteo rápido. El conteo rápido es un procedimiento estadístico para estimar las tendencias de los resultados finales de una elección. Consiste en la selección de una muestra aleatoria estratificada de todas las casillas instaladas el día de la Jornada ELectoral, para con esto estimar el porcentaje de votos para cada uno de los candidatos.

Normalmente, los resultados obtenidos del conteo rápido se hacen públicos utilizando una muestra parcial, ya que la muestra completa tarda en llegar. Se ha observado que rara vez llega la muestra completa, debido a problemas de comunicación en áreas no-urbanas y a condiciones climáticas en ciertas regiones. Dado esto, se sabe que los datos faltantes en la muestra no son aleatorios.

(debería de profundizar mas?)

3.1.2. Definir las observaciones

Una vez que se hizo el análisis conceptual del proceso de medición, se empieza a construir el modelo matemático. Primero se define el espacio en el que las observaciones pueden tomar valores. Para este caso, nuestras observaciones son el número de votos y la lista nominal de cada una de las casillas. El número de votos para una casilla puede tomar valores desde 0 hasta el total de la lista nominal de ésta. La participación, al ser una proporción, puede tomar valores entre 0 y 1. Se buscan distribuciones probabilísticas que cumplan con estas características. La distribución Beta es la principal candidata para modelar proporciones gracias a su versatilidad.

3.1.3. Identificar estadísticas resumen relevantes

Teniendo definido el espacio de las observaciones, se construyen las estadísticas resumen que contienen la información relevante del diseño experimental. Además, con base en dichas estadísitcas se debe pensar qué comportamientos son razonables o no para el fenómeno que se está estudiando.

Para el caso de la participación electoral se puede utilizar como estadísitca resumen la participación electoral total. Después del análisis conceptual anterior, se puede decir que valores por debajo de 30 % y por encima de 90 %, serían considerados como compartamientos

no razonables.

3.2. Modelo

Con el diseño experimental definido, se puede empezar a construir un modelo. Es importante que el modelo tenga definido los siguientes puntos:

- **Consistencia con la experiencia en el dominio de estudio** Para compenzar una función de verosimilitud que no está bien identificada, se necesita una distribución previa que incorpore la suficiente información del dominio para evitar configuraciones extremas del modelo.
- **Cómo se espera que se comporten las inferencias sobre las realizaciones del proceso de medición**
- **Captura de la estructura del proceso generador de los datos**

3.2.1. Construir un modelo

Como ya se sabe, un modelo bayesiano necesita un modelo de las observaciones $\pi_s(y|\theta)$, compuesto por el posible proceso generador de los datos y un modelo a priori $\pi_s(\theta)$ que resume la experiencia en el dominio de estudio.

La idea dentro de este flujo de trabajo bayesiano es que el modelo inicial debe de ser lo más sencillo posible. El modelo debe de ser suficiente para contestar la pregunta de interés, en este ejemplo sería: ¿Cuál es el nivel de participación a nivel nacional?

3.2.1.1. Construir un modelo de las observaciones

El modelo observacional se construye a partir de distribuciones de probabilidad sobre el espacio de observaciones. La colección de distribuciones que definene el proceso generador de los datos debe de ser coherente con lo que se sabe del dominio de estudio. Este modelo es tan solo una aproximación al verdadero proceso de medición, sin embargo, incluso aproximaciones simples pueden ofrecer respuesta a preguntas que surgen en la práctica.

En este ejemplo, la distribución de los votos de cada casilla puede pensarse como una distribución Binomial, ya que una de las interpretaciones de esta distribución es la probabilidad de éxito o fracaso en un experimento que es repetido varias veces. Hablando sobre participación electoral, el éxito sería que un elector ejerza su voto. Los parámetros de la distribución Binomial son n y p, donde n sería la lista nominal de una casilla y p, la probabilidad de que alguien vote dentro de dicha casilla.

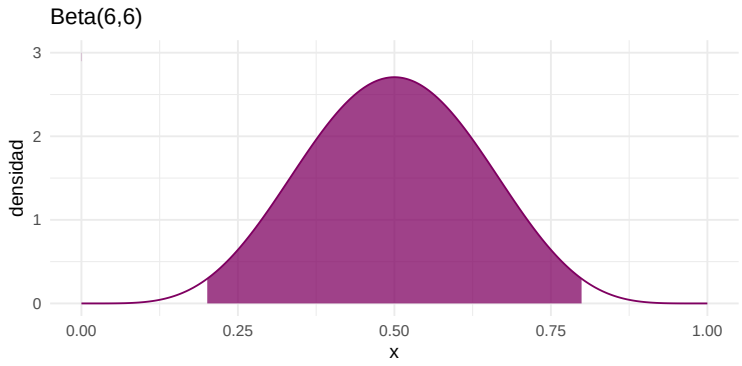
3.2.1.2. Completar el modelo con distibuciones previas

En un modelo bayesiano, el modelo observacional se complementa con una distribución previa sobre el espacio de parámetros $\pi_s(\theta)$. El objetivo de la distribución previa no es que contenga toda la información que se tiene sobre el dominio de estudio, simplemente se espera que el modelo conjunto se comporte adecuadamente...

poner aquí a qué se refiere que se comporte adecuadamente...

En la práctica, es suficiente con que la distribución previa contenga información sobre la escala, unidades u órden de magnitud sobre los parámetros. De lo que se trata es de complementar la información del modelo observacional.

Como se mencionó anteriormente, trabajando con la particiación electoral, un buen candidato para el parámetro p de la distribución Binomial sería la distribución Beta. Se habló de que, históricamente, no se han observado niveles de participación eletoral nacional por debajo del 30 % y por arriba del 90 %. Por lo tanto puede tomarse como previa para la participación nacional una distribución Beta(6,6) que queda centrada en 0.5, es decir, 50 % de participación. La distribución previa se vería como se ilustra a continuación:



Entonces, se tiene un primer modelo para la participación definido por:

- Distribución previa Beta con parámetros de localización y escala iguales a 6 para el nivel de participación nacional (p_{nac})
- La participación de un estrato (p_{est_i}) se distribuye Beta con media p_nac y cierto parámetro de dispersión ϕ
- La participación en cada casilla p_j se distribuye Beta con media $p_{est_{i(j)}}$ y cierto parámetro de dispersión ϕ_i

3.3. Identifcar nuevas estadísticas resumen

3.4. Analizar el ensamble

Incluso antes de que se utilicen los datos observados, se tiene que considerar cuáles son los resutados de los supuestos del modelo.

Este análisis se puede realizar con muestras de las configuraciones del modelo y observaciones de la distribución conjunta $\pi_s(y|\theta)$. Normalmente, primero se simulan configuraciones del modelo de la distribución previa y después del proceso generador de los datos.

Para este caso se van a considerar $R = 1000$ realizaciones de la distribución conjunta, cada una simula los valores observados para las $N =$ casillas.

3.4.1. Analizar la distribución Previa Predictiva

Cada observación simulada del ensamble da un histograma resumen. Si la distribución previa predictiva indica conflicto entre el conocimiento del dominio y el modelo, se tiene que incorporar más información para delimitar de una mejor manera la distribución previa.

3.4.2. Evaluar el ajuste simulado

Además, para cada observación simulada se puede construir una distribución posterior $\pi_s(y|\theta)$. En particular, la distribución construida de las posteriores permite evaluar la exactitud del método computacional.

Dependiendo del método computacional que se utilice existen distintos diagnósticos. Por ejemplo, para métodos de cadenas de Markov Monte Carlo existe la $\hat{R}$.

En este ejemplo se usarán como diagnósticos del ajuste el rango de calibración basada en simulación, el puntaje z (**z-score**) posterior y el "encogimiento" (**shrinkage**) posterior.

- *Calibración basada en simulación*: se compara la muestra del ensamble posterior y la muestra previa usando rangos. Para cada observación simulada se generan R muestras de la distribución posterior correspondiente,

$$\begin{aligned} \tilde{\theta} &\sim \pi_s(\theta) \\ \tilde{y} &\sim \pi_s(y|\tilde{\theta}) \\ (\tilde{\theta}'_1, ..., \tilde{\theta}'_R) &\sim \pi(\theta|\tilde{y}), \end{aligned}$$

y calcular el rango de la muestra previa dentro de las muestras posteriores, es decir, el número de muestras posteriores mayores a la muestra previa,

$$\rho = \#\{\tilde{\theta} < \tilde{\theta}'_r\}.$$

si el ensamble posterior y la muestra previa son equivalentes, entonces los rangos deben de distribuirse de manera uniforme.

3.5. Ajustar las observaciones y evaluar

Una vez que se hicieron los diagnósticos del modelo, se puede construir una distribución posterior con los datos observados.

3.6. Analizar la distribución predictiva posterior

Capítulo 4

Conclusiones

Algunas conclusiones

Apéndice A

Titulo A

Apéndice B

Titulo B

Apéndice C

Notas

Metodología SAE

Definición de los principales indicadores de violencia en la calle. Estimación de áreas pequeñas usando tecnicas *model-assisted* y *model-based*.

Pratesi y Salvati (Monica Pratesi 2016, Capítulo 1) dicen, la estimación de indicadores a un nivel local se calcula con metodos indirectos usando variables auxiliares, usualmente proveniente de datos disponibles a nivel local.

> Considering Särndal *et al.* 1992 we clarify that in this context a model consists of "some assumptions of relationship, unverifiable but not entirely out of place, to save survey resources or to bypass other practical difficulties"

Un modelo ayudará a hacer predicciones de observaciones de la población basado en variables auxiliares. La metodología basada en modelos permitirá la construcción de intervalos de estimadores.

Sin embargo existen varios problemas relacionados a los datos para implementar la metodología. El primero es definiciones incoherentes entre fuentes de información. El segundo, la exactitud de los estimadores a distintos niveles de estimación.

> The keyword is the harmonization of the registers in such a way that information of the registers in such a way that information from different sources and observed data should be consistent and coherent.

Métodos basados en modelos

Generalmente esta metodología, según Pratesi y Salvati (Monica Pratesi 2016, Capítulo 1), lo más común es usar estimación con modelos lineales mixtos.

La ecuación C.1 presenta el modelo generalizado.

$$y_{jd} = \boldsymbol{x}_{jd}^{\boldsymbol{T}}\boldsymbol{\beta} + u_d + e_{jd} \tag{C.1}$$

donde $\boldsymbol{x}_{jd}$ representa el conjunto de variables auxiliares, u_d representa el efecto aleatorio de area y e_{jd} representa el efecto aleatorio individual.

C.1. Target Parameters

C.2. Ejemplos de citas, referencias

Un ejemplo de cita es (R Core Team 2016).

Otro ejemplo . Blah blah

Un ejemplo de cita es (R Core Team 2016; Wickham 2019).

$$f\left(k\right) = \binom{n}{k} p^k \left(1-p\right)^{n-k} \tag{C.2}$$

You may refer to it using C.2, e.g., see Equation

En la sección 1.

Referencias

Monica Pratesi, ed. 2016. *Analysis of Poverty Data by Small Area Estimation*. 2nd ed. Wiley.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2019. *stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Índice alfabético