EST-46115: Modelación Bayesiana

Profesor: Alfredo Garbuno Iñigo — Primavera, 2023.
Objetivo: Repasar/Introducir notación que utilizaremos a lo largo del curso. Y a la vez, establecer la motivación de los temas que trataremos en la materia.
Lecturas recomendas: Notas del curso de fundamentos (2021) y sección 1 de [2].

1. MOTIVACIÓN

Cualquier tarea de modelado basado en datos está sujeta a incertidumbre. Como científicos de datos, tendrán que tomar o informar decisiones basándose en la información disponible. Por lo tanto, es natural que tengan que incorporar incertidumbre en sus análisis.

Como científicos aplicados lo que desean hacer es aproximar un proceso físico (tangible) por medio de modelos matemáticos.

La precisión con la que nuestro modelo puede aproximar la realidad, esto es la diferencia o `la discrepancia` entre modelo y realidad, se debe a la `incertidumbre` inherente en nuestro proceso de modelado. Dicha incertidumbre la podemos considerar como consecuencia de dos tipos:

1. `Incertidumbre aleatoria`: también conocida como incertidumbre estadística, estocástica o irreducible. Se refiere a la incertidumbre que es natural para nuestro proceso y que no podemos reducir por medio de un mejor modelo.
2. `Incertidumbre epistémica`: se refiere a la incertidumbre derivada de nuestra simplificación del problema, nuestro estado de conocimiento o supuestos. En algunas ocasiones está asociada a los métodos numéricos con los que implementamos nuestros modelos. En otras, está asociada con los supuestos con lo que contamos para resolver un problema.

Esta distinción nos ayuda a visualizar dos conceptos:

1. Identificar la necesidad de modelar incertidumbre en nuestros procesos.
2. Identificar el origen de dicha incertidumbre.

Lamentablemente en la práctica, al momento de generar simulaciones, nos olvidamos estas nociones y siempre es importante considerar las limitaciones de nuestros modelos para representar correctamente el proceso que nos interesa.

Ahora, la pregunta natural es ¿cómo modelamos la incertidumbre? En este curso (y en general en cualquier otras aplicaciones) utilizaremos el `lenguaje de probabilidad` para `expresar incertidumbre` ([3]). En este enfoque, es usual considerar bajo nuestros procesos de inferencia incertidumbre aleatoria y epistémica.

En un proceso de modelado completo la incertidumbre puede deberse a distintos factores. Es usual seguir abstraer este procedimiento por medio del siguiente par de ecuaciones

$$z = y + \epsilon\,, \tag{1}$$
$$y = f(x) + \varepsilon\,. \tag{2}$$

El curso busca desmitificar la noción de incorporar incertidumbre en nuestro proceso de modelado. Esto es por que:

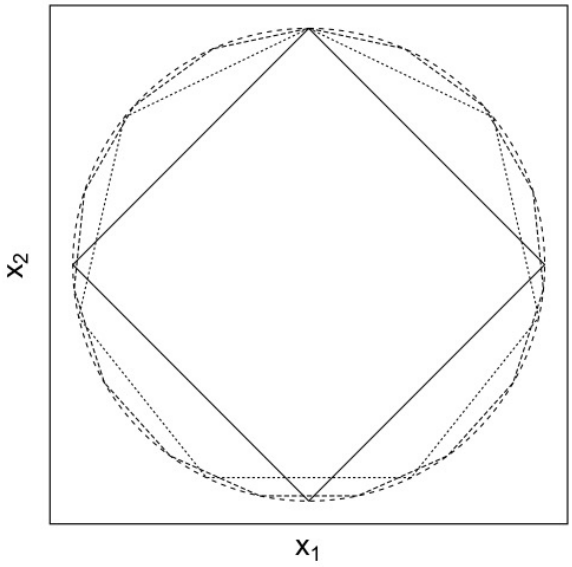


FIGURA 1. *Aproximación a un circulo mediante una trayectoria discretizada.*

1. Hay una falso sentido de seguridad por llamar cualquier modelo `bayesiano`;
2. El uso de `herramientas computacionales automáticas` nos puede hacer caer en el modelado bajo cajas negras.

Consideremos el conjunto de datos siguiente.

```
data(mpg)
data ← mpg ▷ as_tibble()
data ▷ print(n = 5)
```

```
# A tibble: 234 × 11
  manufacturer model displ  year   cyl trans      drv      cty   hwy fl
     class
  <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr
     >
1 audi         a4      1.8  1999     4 auto(15)   f         18    29 p ...
    compa
2 audi         a4      1.8  1999     4 manual(m5) f         21    29 p ...
    compa
3 audi         a4      2    2008     4 manual(m6) f         20    31 p ...
    compa
4 audi         a4      2    2008     4 auto(av)   f         21    30 p ...
    compa
5 audi         a4      2.8  1999     6 auto(15)   f         16    26 p ...
    compa
# ... with 229 more rows
#  Use ‘print(n = ...)‘ to see more rows
```

Nos interesa poder relacionar el rendimiento de un auto en carretera y el rendimiento del mismo en una ciudad, ver Fig. 2. Operativamente lo podemos realizar con los siguientes comandos.

```
glm(hwy ∼ cty, data, family = gaussian()) ▷
```

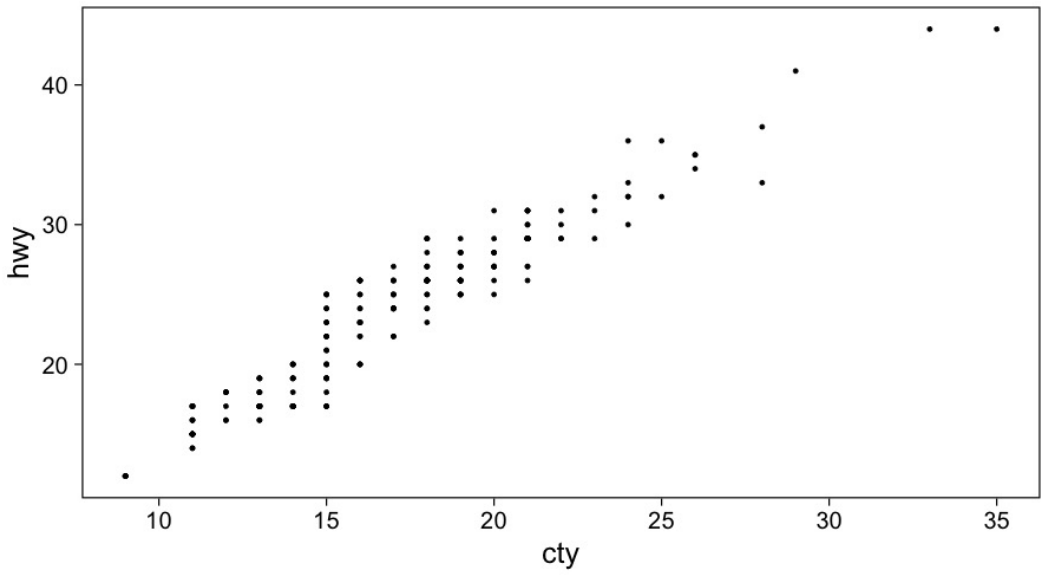


FIGURA 2. *Rendimiento en carretera y rendimiento en ciudad.*

```
  summary()
```

```

Call:
glm(formula = hwy ~ cty, family = gaussian(), data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-5.341  -1.279   0.021   1.034   4.046

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.892      0.469     1.9    0.058 .
cty            1.337      0.027    49.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1

(Dispersion parameter for gaussian family taken to be 3.071)

    Null deviance: 8261.66  on 233  degrees of freedom
Residual deviance:  712.36  on 232  degrees of freedom
AIC: 930.6

Number of Fisher Scoring iterations: 2
```

```
library(rstanarm)
stan_glm(hwy ~ cty, data = data, refresh = 0) ▷
  summary()
```

```

Model Info:
```

```
 function:     stan_glm
 family:       gaussian [identity]
 formula:      hwy ~ cty
 algorithm:    sampling
 sample:       4000 (posterior sample size)
 priors:       see help('prior_summary')
 observations: 234
 predictors:   2

Estimates:
              mean   sd   10%   50%   90%
(Intercept) 0.9    0.5  0.3   0.9   1.5
cty         1.3    0.0  1.3   1.3   1.4
sigma       1.8    0.1  1.7   1.8   1.9

Fit Diagnostics:
           mean   sd   10%   50%   90%
mean_PPD 23.4    0.2 23.2  23.4  23.6

The mean_ppd is the sample average posterior predictive distribution of the
    outcome variable (for details see help('summary.stanreg')).

MCMC diagnostics
              mcse Rhat n_eff
(Intercept)   0.0  1.0  4003
cty           0.0  1.0  4107
sigma         0.0  1.0  3829
mean_PPD      0.0  1.0  3775
log-posterior 0.0  1.0  1859

For each parameter, mcse is Monte Carlo standard error, n_eff is a crude
    measure of effective sample size, and Rhat is the potential scale
    reduction factor on split chains (at convergence Rhat=1).
```

Los resúmenes de ambos modelos son similares. A simple vista parece que sólo cambio la forma de ajustar un modelo en lugar de `glm` utilizamos la función `rstanarm::stan_glm`. ¿Qué es lo que cambia?

2. NOTACIÓN

Denotamos por x una `variable aleatoria` y por $\mathbb{P}(\cdot)$ una `función de distribución`. Escribimos $x \sim \mathbb{P}$ para denotar que la variable aleatoria x tiene distribución $\mathbb{P}(\cdot)$. Denotamos por $\mathbb{E}[\cdot]$ el `valor esperado` del argumento con respecto a la distribución que estamos considerando. Durante el curso seremos explícitos en la variable aleatoria y usaremos

$$\mathbb{E}_x[\cdot] = \int_{\mathcal{X}} \cdot \, \pi(x) \, \mathrm{d}x \,, \tag{3}$$

o bien, haremos énfasis en la distribución por medio de lo siguiente

$$\mathbb{E}_\pi[\cdot] = \int_{\mathcal{X}} \cdot \, \pi(x) \, \mathrm{d}x \,, \tag{4}$$

de acuerdo al contexto.

> Nota que en las ecuaciones anteriores estamos considerando el término $\pi(\cdot)$ como la `función de densidad` de la `función de probabilidad` $\mathbb{P}(\cdot)$.

Nos será útil la siguiente notación para evaluar valores esperados

$$\pi(f) := \mathbb{E}_\pi[f(x)] = \int_{\mathcal{X}} f(x)\,\pi(x)\,\mathrm{d}x\,, \tag{5}$$

pues será el `objetivo general` para los métodos que estudiaremos en el curso.

Por ejemplo, utilizaremos la noción de `aproximar integrales` por medio de algún procedimiento de muestreo de tal forma que esperaremos construir un estimación $\hat{\pi}(f)$ con cierto grado de refinamiento. Por ejemplo, veremos el `método Monte Carlo` que utiliza una colección de N simulaciones para aproximar la integral anterior. Esto lo denotaremos por

$$\hat{\pi}_N(f) \approx \pi(f)\,. \tag{6}$$

En general, nos interesa, y esperamos que, podamos:

1. Mejorar nuestra estimación con mas simulaciones

$$\lim_{N\to\infty} \hat{\pi}_N(f) = \pi(f) \tag{7}$$

2. Cuantificar la incertidumbre en nuestra aproximación por medio de alguna distribución de probabilidad. Por ejemplo,

$$\hat{\pi}_N(f) \sim \mathsf{N}\left(\pi(f), \frac{\mathbb{V}(f)}{N}\right)\,. \tag{8}$$

2.0.1. ***Definición*** *[Distribución paramétrica]:* Decimos que una función de distribución es `paramétrica` si se puede identificar completamente la distribución con respecto a un `vector de parámetros` $\theta \in \mathbb{R}^p$. Esto lo denotamos de la siguiente manera

$$\pi_\theta(x) \qquad \text{ó} \qquad \pi(x;\theta)\,, \tag{9}$$

y si $\theta \neq \theta'$ entonces $\pi_\theta(x) \neq \pi_{\theta'}(x)$ para cualquier x en el `soporte`.

3. REPASO DE PROBABILIDAD

Consideraremos como requisitos el contenido de `Fundamentos de estadística` o equivalentes. En particular lo que requerimos como base es lo siguiente.

3.0.1. ***Definición [Espacio de Probabilidad]:*** Un espacio de probabilidad está definido por la terna $(\Omega, \mathcal{X}, \mathbb{P})$:

1. El espacio muestral, Ω (elementos).
2. El espacio de eventos medibles, $\mathcal{X}$ (subconjuntos).
3. La medida de probabilidad, $\mathbb{P} : \mathcal{X} \to [0,1]$.

3.0.2. ***Definición [Variable aleatoria]:*** Una variable aleatoria es una función $X : \mathcal{X} \to \mathbb{R}$ con la propiedad de que las pre-imágenes bajo X son eventos medibles. Es decir,

$$\{w \in \mathcal{X} : X(w) \leq x\} \in \mathcal{X} \qquad \forall x \in \mathbb{R}. \tag{10}$$

3.0.3. ***Definición [Función de acumulación]:*** Para toda variable aleatoria X tenemos una función de acumulación $\mathbb{P}_X : \mathbb{R} \to [0,1]$ dada por

$$\mathbb{P}_X(x) = \mathbb{P}\big(\{w \in \mathcal{X} : X(w) \leq x\}\big)\,. \tag{11}$$

Esto usualmente lo escribimos como $\mathbb{P}_X(x) = \mathbb{P}\{X \leq x\}$.

3.0.4. ***Definición [Función de densidad]:*** Una variable aleatoria es continua si su función de acumulación es `absolutamente continua` y puede ser expresada por medio de

$$\mathbb{P}_X(x) = \int_{-\infty}^{x} \pi(s)\,\mathrm{d}s\,, \tag{12}$$

donde la anti-derivada $\pi : \mathbb{R} \to [0, \infty)$ se llama la `función de densidad` de la variable aleatoria X.

Las propiedades generales de las distribuciones de probabilidad se pueden especificar por medio de su centralidad (localización), su dispersión, su rango de valores, su simetría y el comportamiento de valores extremos.

En general esto lo podemos extraer de los momentos

$$\mathbb{E}(X^p) = \int_{\mathbb{R}} x^p\,\pi(x)\,\mathrm{d}x\,, \tag{13}$$

o los momentos centrales. Por ejemplo: media y varianza.

Uno de los resultados que espero recuerden bien de sus cursos anteriores es el de la `Ley de los Grandes Números`. La cual podemos enunciar como:

3.0.5. ***Teorema [Ley de los Grandes Números]:*** Sea $X_1, X_2, \ldots$ una colección de variables aleatorias independientes e idénticamente distribuidas (iid) y sea $\bar{X}_n$ el promedio de un subconjunto de n. Si denotamos por μ el valor promedio de X_i dentro de esa colección, entonces tenemos que

$$\bar{X}_n \to \mu \quad \text{(casi seguramente)}\,. \tag{14}$$

3.0.6. ***Teorema [Límite Central]:*** Sea $X_1, \ldots, X_n$ una colección de n variables aleatorias iid con $\mathbb{E}[X_i] = \mu$ y $\mathbb{V}[X_i] = \sigma^2 < \infty$. Entonces

$$\bar{X}_n \sim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right)\,, \tag{15}$$

para n suficientemente grande.

3.0.7. Para pensar ¿Qué es probabilidad?

4. REPASO INFERENCIA

Repaso de inferencia bajo un enfoque frecuentista.

4.1. Regla de Bayes

La `regla de Bayes` utiliza la definición de probabilidad condicional para hacer inferencia a través de

$$\pi(A|B) = \frac{\pi(B|A)\pi(A)}{\pi(B)}\,. \tag{16}$$

4.2. Ejemplos

- Verosimilitud: $x|\theta \sim \mathsf{Binomial}(n, \theta)$ + Previa: $\theta \sim \mathsf{Beta}(\alpha, \beta)$ = Posterior: ?
- Verosimilitud: $x|\theta \sim \mathsf{Uniforme}(0, \theta)$ + Previa: $\theta \sim \mathsf{Pareto}(\alpha, \theta_0)$ = Posterior: ?

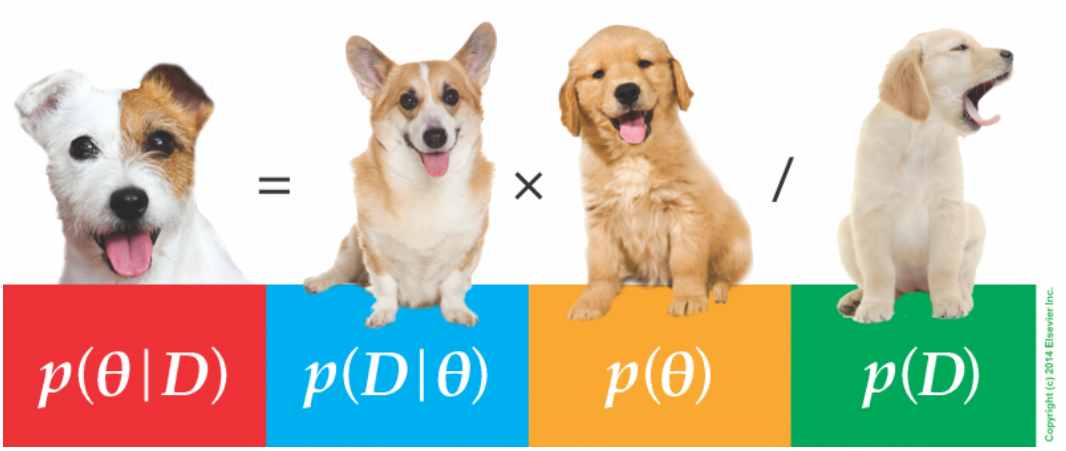


FIGURA 3. *Tomado de [5]* .

5. REPASO INFERENCIA

Repaso de inferencia bajo un enfoque bayesiano.

5.1. Ejemplo

Este ejemplo fue tomado de [4].

5.2. Diferentes previas, diferentes posteriores

```
modelo_beta ← function(params, n = 5000){
  rbeta(n, params$alpha, params$beta)
}
```

```
escenarios ←
  tibble(analista = fct_inorder(c("Ignorante", "Indiferente",
                                  "Feminista", "Ingenuo")),
         alpha = c(1, .5, 5, 14),
         beta  = c(1, .5, 11, 1)) ▷
  nest(params.previa = c(alpha, beta)) ▷
  mutate(muestras.previa = map(params.previa, modelo_beta))
```

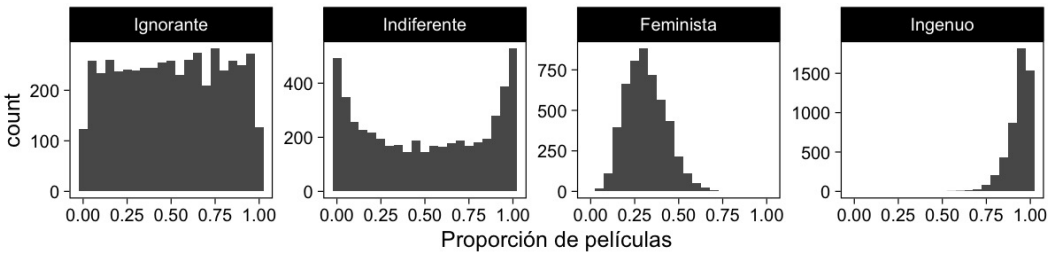


FIGURA 4. *Muestras de* $\theta \sim$ Previa .

```
update_rule ← function(params){
```

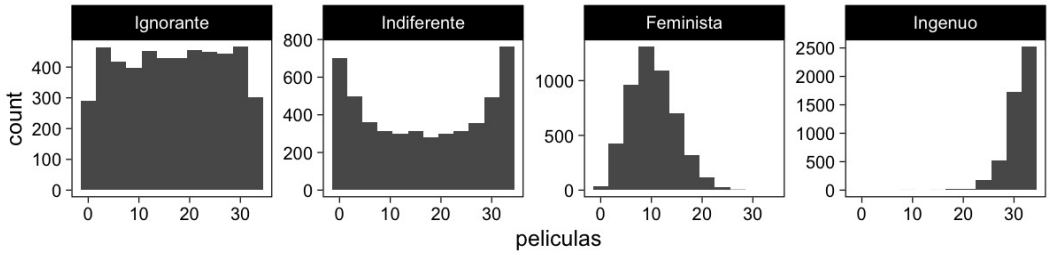


FIGURA 5. *Distribución predictiva previa*

```
  tibble(alpha = params$alpha + data$PASS,
         beta  = params$beta  + data$FAIL)
}
escenarios ← escenarios ▷
  mutate(params.posterior = map(params.previa, update_rule),
         muestras.posterior = map(params.posterior, modelo_beta))
```

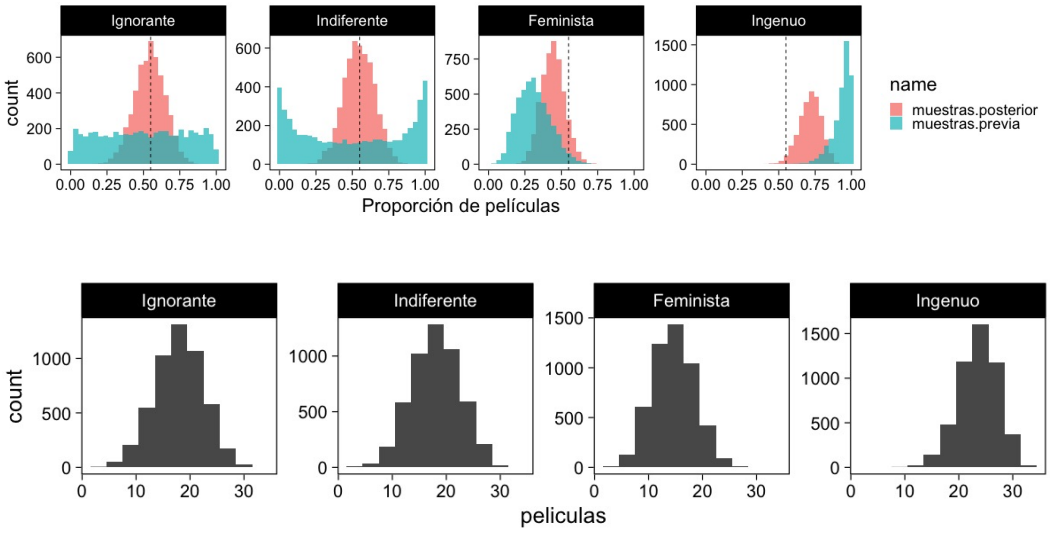
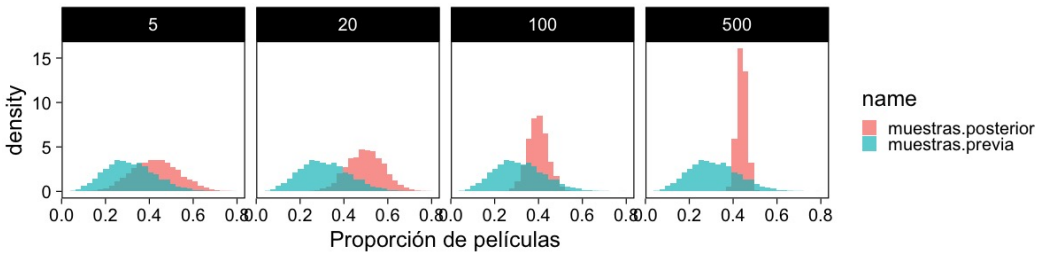


FIGURA 6. *Predictiva posterior.*

5.3. Diferentes datos, diferentes posteriores



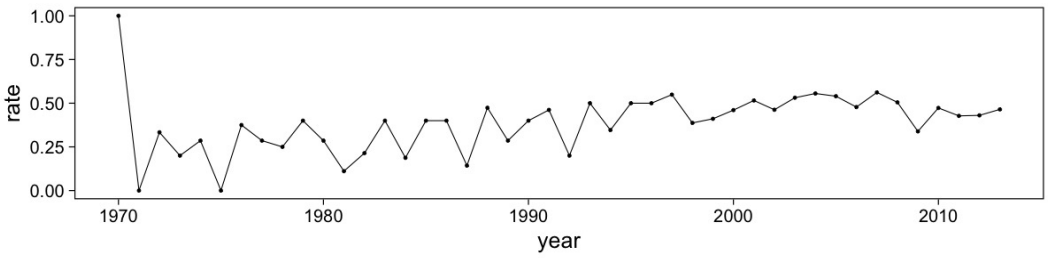


FIGURA 7. *Histórico de la proporción de peliculas que pasan la prueba de Bechdel por año.*

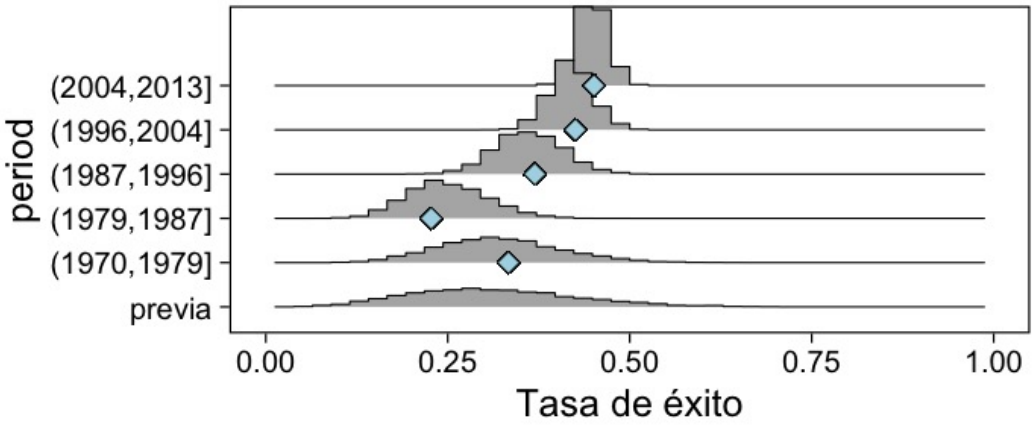


FIGURA 8. *La posterior de hoy puede ser la previa de mañana.*

5.4. Análisis secuencial

5.5. Tarea

Echenle un ojo a la sección 5.2 de Bayes rules! donde se expone a detalle un modelo más del análisis conjugado. ¿Puedes identificar/derivar la distribución predictiva?

6. ¿QUÉ VEREMOS?

Por medio de metodología Bayesiana podemos cuantificar incertidumbre en:

- Observaciones.
- Parámetros.
- Estructura.

Es fácil especificar y ajustar modelos. Pero hay preguntas cuyas respuestas no han quedado claras:

1. Construcción.
2. Evaluación.
3. Uso.

> Programación probabilística.

Los aspectos del flujo de trabajo Bayesiano consideran ([2]):

1. Construcción iterativa de modelos.

2. Validación de modelo (computacional).
3. Entendimiento de modelo.
4. Evaluación de modelo.

6.1. Distinción importante

`Inferencia` no es lo mismo que `análisis de datos` o que un `flujo de trabajo`.

> Inferencia (en el contexto bayesiano) es formular y calcular con probabilidades condicionales.

6.2. ¿Por qué necesitamos un flujo de trabajo?

- El cómputo puede ser complejo.
- Expandir nuestro entendimiento en aplicaciones.
- Entender la relación entre modelos.
- Distintos modelos pueden llegar a distintas conclusiones.

6.3. Proceso iterativo

- La gente de ML sabe que el proceso de construcción de un modelo es iterativo, ¿por qué no utilizarlo?

> Una posible explicación puede encontrarse en [1]. El argumento es formal en cuanto a actualizar nuestras creencias como bayesianos. Sin embargo, con cuidado y un procedimiento científico puede resolver el asunto.

REFERENCIAS

[1] A. Gelman and Y. Yao. Holes in Bayesian statistics. *Journal of Physics G: Nuclear and Particle Physics*, 48(1):014002, jan 2021. ISSN 0954-3899, 1361-6471. . 10

[2] A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020. 1, 9, 11

[3] E. Jaynes and G. Bretthorst. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. ISBN 978-0-521-59271-0. 1

[4] A. Johnson, M. Ott, and M. Dogucu. *Bayes Rules! An Introduction to Applied Bayesian Modeling*. 2021. 7

[5] J. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, 2014. 7

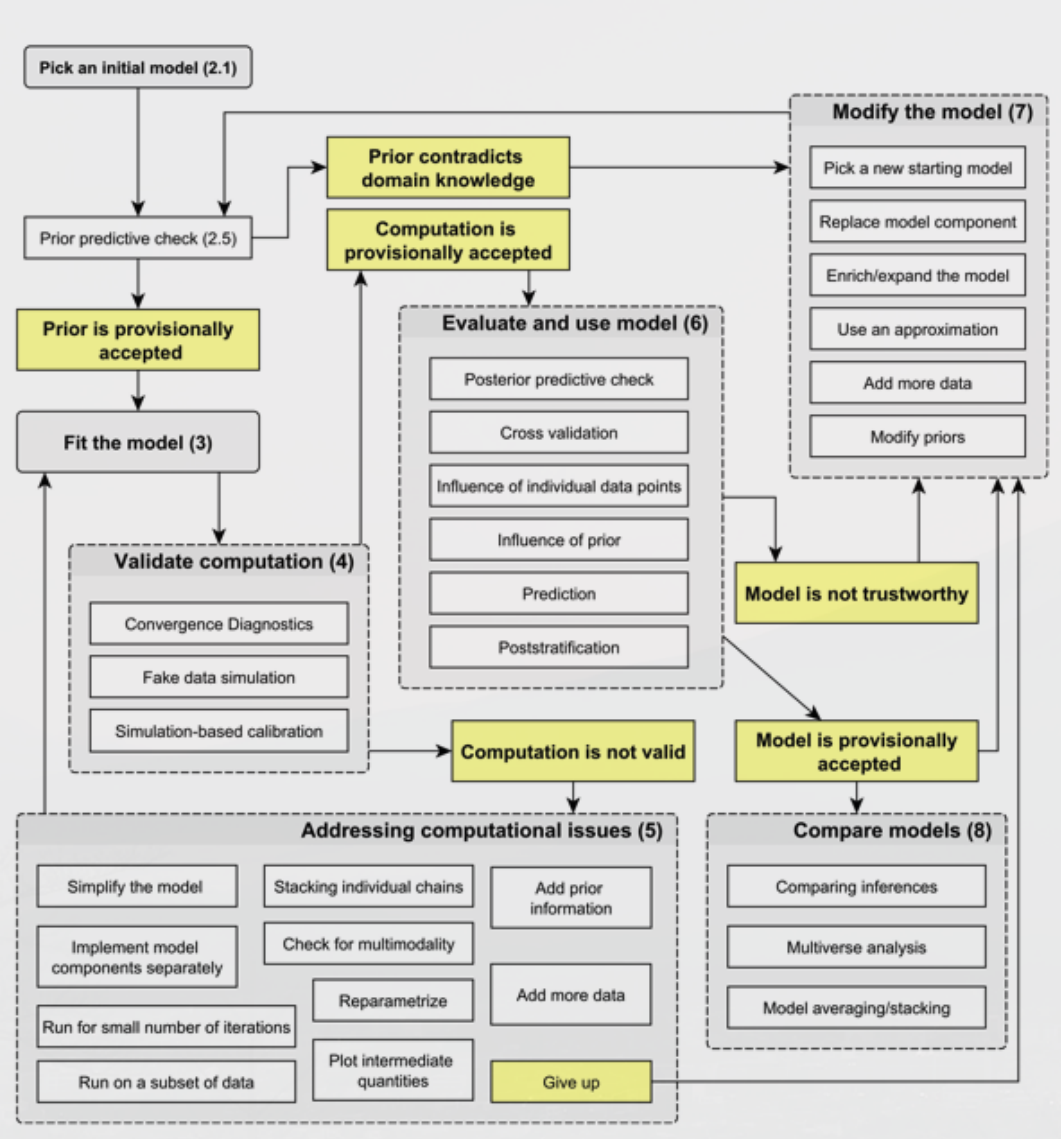


FIGURA 9. *Tomado de [2].*