

El proyecto Manhattan y la estimación desagregada con encuestas de hogares

Diferentes enfoques de CEPAL en el mapeo de la pobreza

Andrés Gutiérrez

2021

¿Por qué hacemos lo que hacemos?

ODS



Poner fin a la pobreza

1

- 1.1. De aquí a 2030, erradicar para todas las personas y en todo el mundo la pobreza extrema.
- 1.2. De aquí a 2030, reducir al menos a la mitad la proporción de hombres, mujeres y niños de todas las edades que viven en la pobreza en todas sus dimensiones con arreglo a las definiciones nacionales.

No dejar a nadie atrás



Desglosar los ODS por ingreso, sexo, edad, raza, etnicidad, estado migratorio, discapacidad y ubicación geográfica, de conformidad con los Principios Fundamentales de las Estadísticas Oficiales.

Marco de indicadores globales para los Objetivos de Desarrollo Sostenible (A/RES/71/313).

Limitaciones de las encuestas de hogares y el uso de información auxiliar

¿De qué se trata?

Las encuestas que dependen de un buen tamaño de muestra y una estrategia de muestreo adecuada se basan en un sistema inferencial sólido que proporciona una estimación precisa y exacta en los dominios planificados.

Cuando el tamaño muestral de la encuesta no es suficiente para sustentar la inferencia estadística requerida para algunos subgrupos de interés, es necesario recurrir a información auxiliar externa para que en conjunto se puede construir un sistema inferencial preciso y exacto.

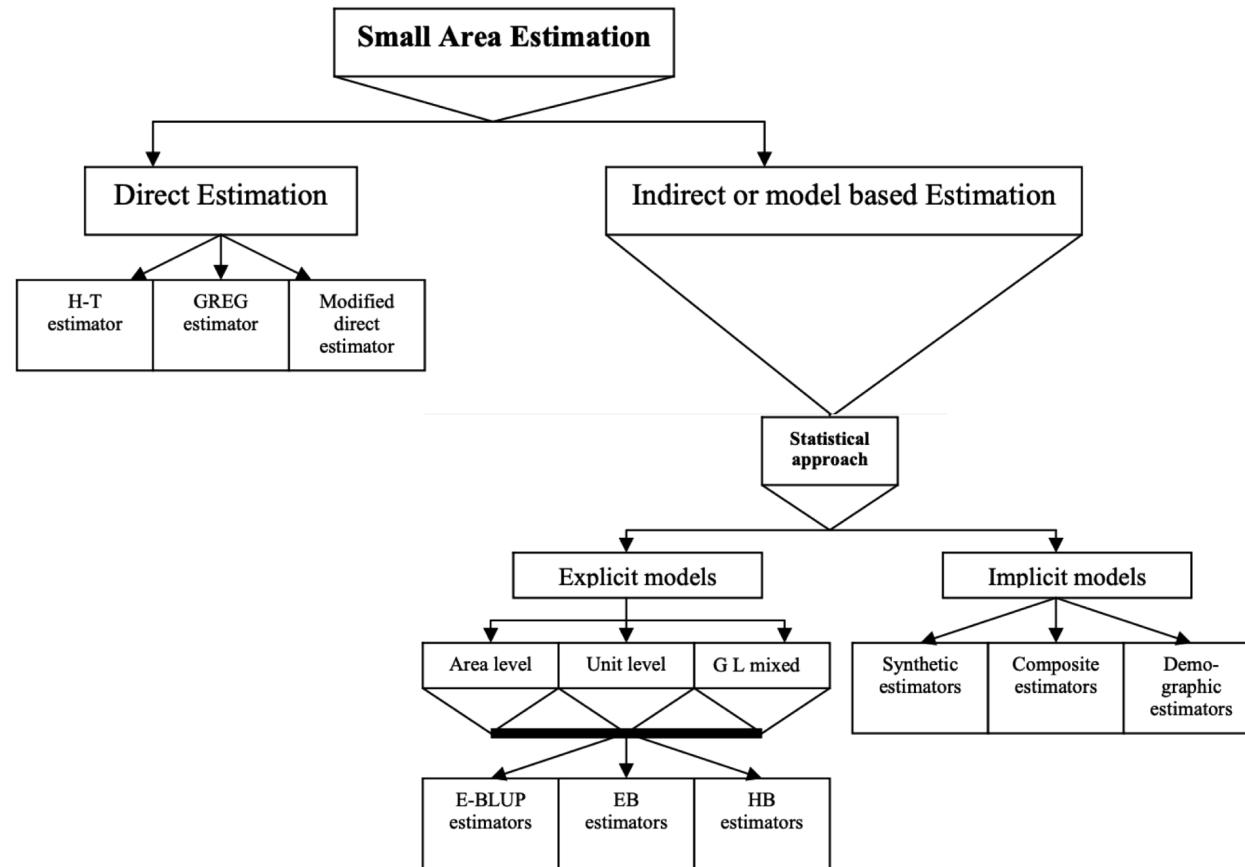
Soluciones

Cuando el tamaño de la muestra no permite obtener estimaciones directas confiables para algunos dominios de interés, se pueden abordar las siguientes opciones:

1. Incrementar el tamaño de la muestra: esta opción eleva los costos y es inviable.
2. Utilizar metodologías estadísticas que involucren información auxiliar externa para obtener estimaciones confiables (no directas) en los subgrupos de interés, mientras se mantiene el tamaño de la muestra de la encuesta.

Metodologías SAE en la CEPAL

Metodologías SAE



Source: adaptation from Rahman (2008).

Dos clases de métodos

Los estimadores de SAE se dividen en dos tipos principales:

1. Estimadores basados en modelos de área
2. Estimadores basados en modelos de unidad

La escogencia del método que se debe utilizar en la estimación de los dominios de interés se realiza dependiendo del nivel en el que se encuentre la información auxiliar (a nivel de dominio o agregación - a nivel de hogar o persona).

Estimadores directos

- El tamaño de muestra en cada área difícilmente es planificado de antemano (pues el esquema de muestreo es bietápico: UPM - Vivienda).
- Cualquier estimación de indicadores relativos (medias, proporciones) tendrá que usar un estimador de razón: *numerador y denominador aleatorios*.

Cuando el tamaño de muestra no es lo suficientemente grande, entonces ninguno de los anteriores estimadores será preciso, ni consistente.

Modelos de áreas con el enfoque de *Tom*

Y te levantas un día...

Y te sientes un poco raro, y débil. Vas al médico y te hacen exámenes. Uno de ellos te marca positivo para una enfermedad muy rara que solo afecta al 0.1% de la población.

No son buenas noticias.

Vas al consultorio del médico y le preguntas qué tan específico es el examen. Te dice que es muy preciso; identifica correctamente al 99% de la gente que tiene la enfermedad.

Siguen siendo malas noticias

Y aún en negación te preguntas: **¿cuál es la probabilidad de que tenga esa enfermedad?**

Y conoces a Thomas...

Esta es la información que tienes:

- $P(E) = 0.001$
- $P(+|E) = 0.99$
- $P(-E) = 0.999$
- $P(+|-E) = 0.01$

Además, por el teorema de probabilidad total

$$\begin{aligned}P(+) &= P(E)P(+|E) + P(-E)P(+|-E) \\&= 0.001 * 0.99 + 0.999 * 0.01 \\&= 0.01098\end{aligned}$$

La regla de Bayes afirma lo siguiente:

$$Pr(E|+) = \frac{Pr(+|E) \times Pr(E)}{Pr(+)}$$

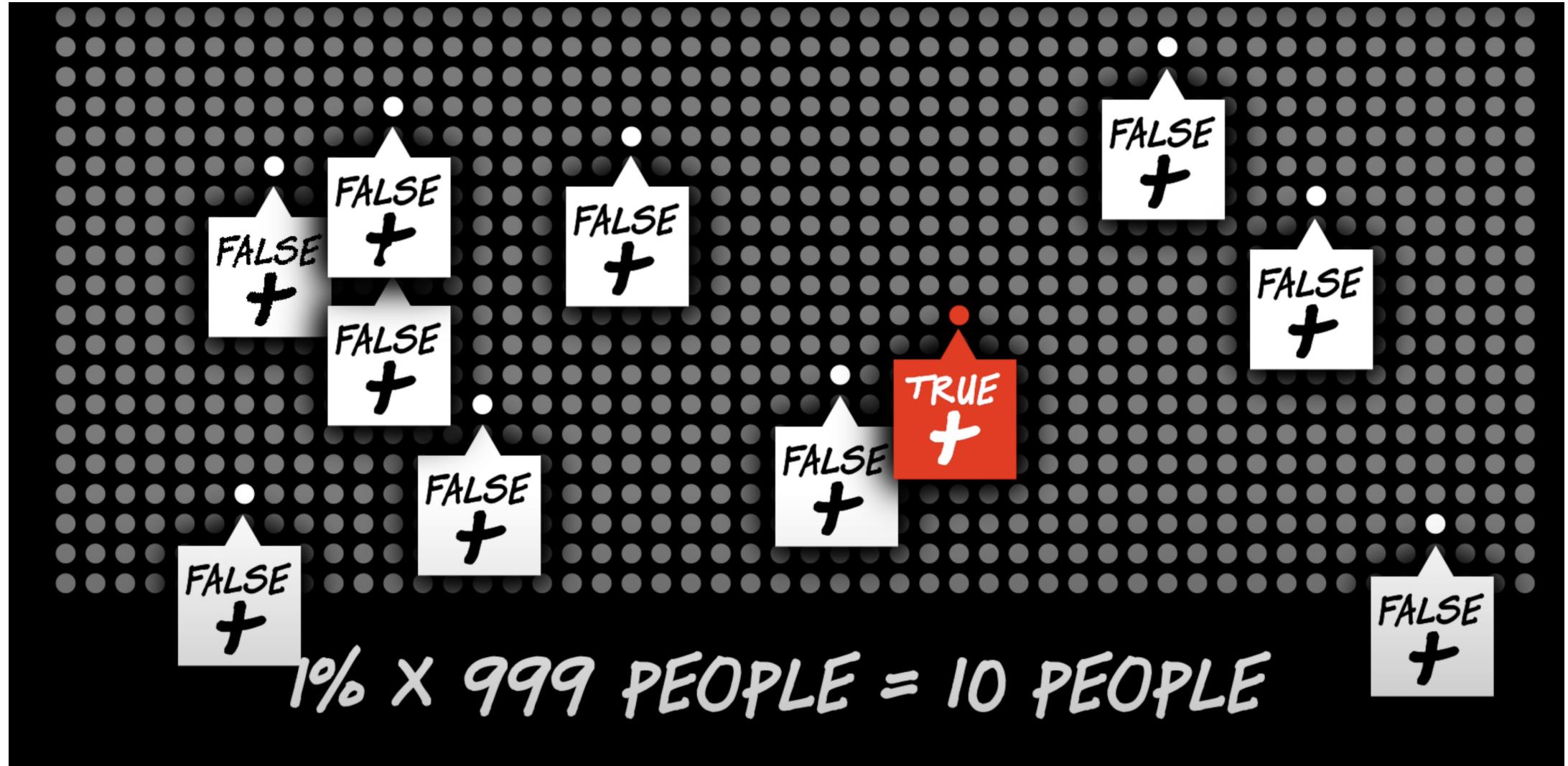
Por lo tanto:

$$Pr(E|+) = 0.09 \approx 9\%$$



icanhasGIF.com

¿Cómo funciona?



Tomado de: [The Bayesian Trap](#)

¿Cómo funciona?



1 IN 11 PEOPLE = 9%

Y pides una segunda opinión

Y esta vez el médico ordena que vuelves a realizarte ese mismo examen... y vuelves a marcar positivo para esa enfermedad.

Y vuelves a preguntarte: **¿cuál es la probabilidad de que tenga esa enfermedad?**

Esta vez, has actualizado tu información sobre $\Pr(E)$, pues ya marcaste positivo en un examen.

$$\Pr(E) = 0.09 \quad \text{y} \quad \Pr(-E) = 0.91$$

Por lo tanto:

$$\Pr(E|++) = 0.907 \approx 91\%$$

El modelamiento bayesiano

En general un modelo bayesiano transforma la información que se tiene acerca del fenómeno en distribuciones de probabilidad debe tener los siguientes elementos:

1. Verosimilitud: $f(y|\theta)$ distribución asociada a los datos observados.
2. Distribución previa: $f(\theta)$ distribución asociada a la incertidumbre de la localización de los parámetros del modelo.

Al aplicar la regla de Bayes, se tiene que

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)}$$

Como $f(y)$ no depende de θ , entonces solo nos preocupamos por el numerador; es decir:

$$f(\theta|y) \propto f(y|\theta)f(\theta)$$

Ejemplo 1

$$y_1, \dots, y_n | \theta \sim N(\theta, 1)$$

$$\theta \sim N(0, 100^2)$$

$$\theta | y \propto N(\theta, \sigma_y^2) \times N(0, 100^2)$$

$$\propto N\left(\bar{y}, \frac{\sigma_y^2}{n}\right)$$

Ejemplo 2

$$y_1, \dots, y_n | \theta \sim N(\theta, \sigma_y^2)$$

$$\theta \sim N(\mu_0, \sigma_0^2)$$

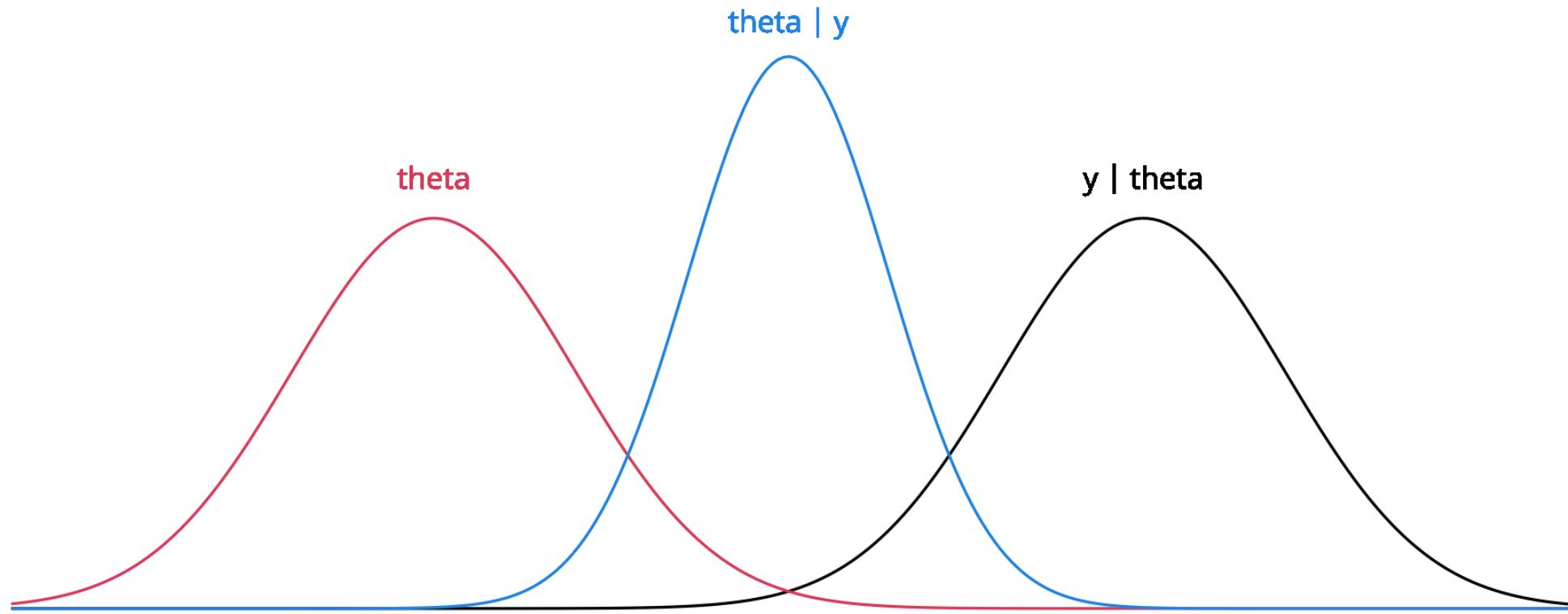
$$\theta | y \propto N(\theta, \sigma_y^2) \times N(\mu_0, \sigma_0^2)$$

$$\propto N(\mu_n, \sigma_n^2)$$

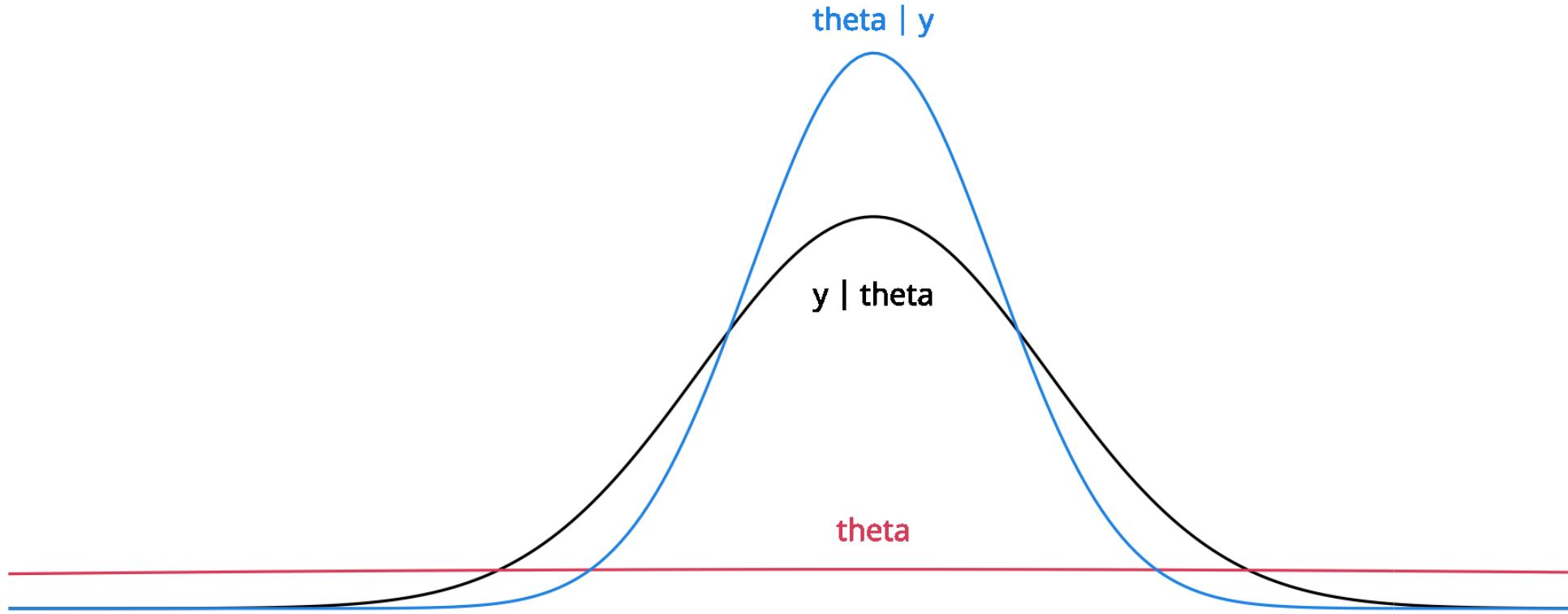
$$\mu_n = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n}{\sigma_y^2} \bar{y}}{\frac{\mu_0}{\sigma_0^2} + \frac{n}{\sigma_y^2}}$$

$$\sigma_n^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma_0^2} \right)$$

Distribución previa informativa sobre θ



Distribución previa no informativa sobre θ



El modelo de Fay-Herriot (I)

A partir de la información en las encuestas tenemos:

1. Un estimador directo para la proporción de personas pobres en la comuna $\hat{\theta}_d^{\text{Dir}}$.
2. Un estimador directo de su varianza $\widehat{\text{Var}}(\hat{\theta}_d^{\text{Dir}})$.

Pensemos un poco en estos estimadores

- El tamaño de muestra en cada comuna difícilmente es planificado de antemano (pues el esquema de muestreo es bietálico: UPM - Vivienda).
- Cualquier estimación de indicadores relativos (medias, proporciones) tendrá que utilizar un estimador de razón: *numerador y denominador aleatorios*.

$$\hat{\theta}_d^{\text{Dir}} = \frac{\sum_{s_d} w_k y_k}{\sum_{s_d} w_k} \quad \widehat{\text{Var}}(\hat{\theta}_d^{\text{Dir}}) = \widehat{\text{AV}}(\hat{\theta}_d^{\text{Dir}}) = \sum_s \sum_k \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}$$

Cuando el tamaño de muestra $n_d = \#(s_d)$ no es lo suficientemente grande, entonces ninguno de los anteriores estimadores será insesgado, ni consistente.

El modelo de Fay-Herriot (II)

Por tanto vamos a modelar el indicador directo para que en las áreas en las que no haya suficiente muestra se tome fuerza prestada de las otras áreas.

$$\begin{aligned}\hat{\theta}_d^{\text{Dir}} &= \theta_d + \varepsilon_d & \varepsilon_d &\sim N(0, \text{Var}(\hat{\theta}_d^{\text{Dir}})) \\ \theta_d &= \mathbf{x}'_d \boldsymbol{\beta} + u_d & u_d &\sim N(0, \sigma_u^2)\end{aligned}$$

$$\hat{\theta}_d^{\text{Dir}} = \mathbf{x}'_d \boldsymbol{\beta} + u_d + \varepsilon_d$$

Aplicando la regla de Bayes, tenemos que

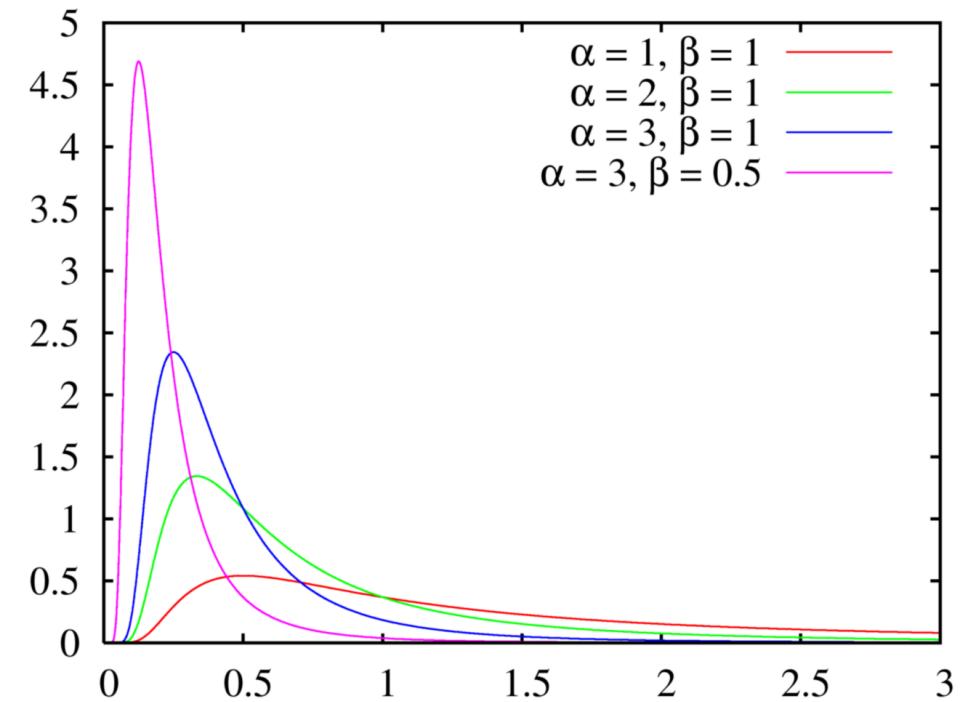
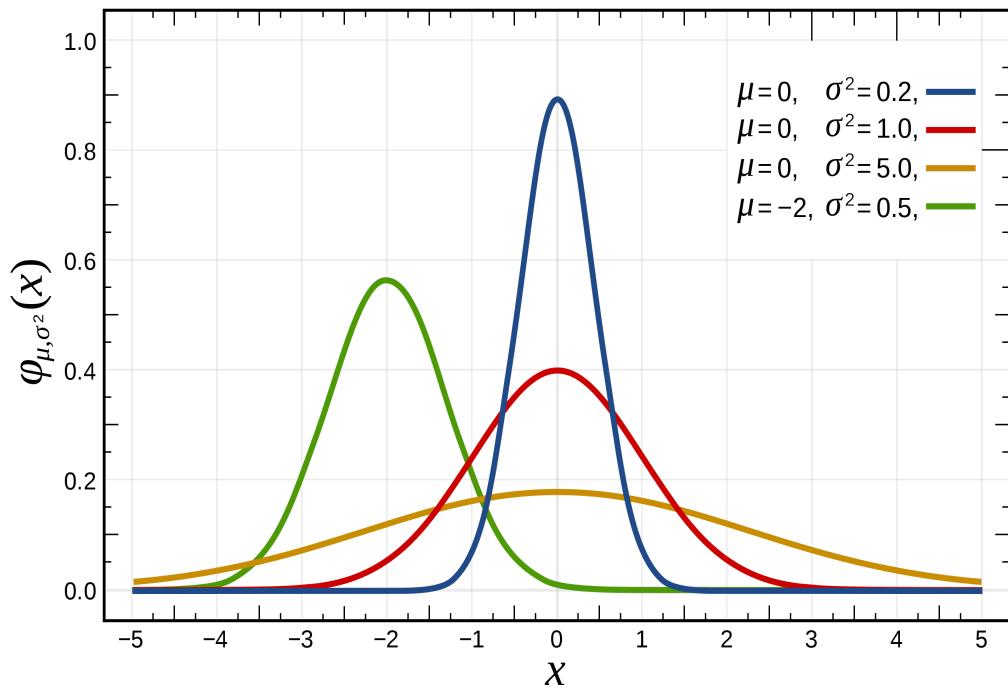
$$\theta_d | \hat{\theta}_d^{\text{Dir}} \sim N\left(\theta_d^{\text{FH}}, \sigma_{d_{\text{FH}}}^2\right)$$

$$\begin{aligned}\theta_d^{\text{FH}} &= E(\theta_d | \hat{\theta}_d^{\text{Dir}}) = \gamma_d \hat{\theta}_d^{\text{Dir}} + (1 - \gamma_d) \mathbf{x}'_d \boldsymbol{\beta} \\ \sigma_{d_{\text{FH}}}^2 &= \text{Var}(\hat{\theta}_d^{\text{Dir}}) \gamma_d\end{aligned}$$

Modelos bayesianos de área

Distribuciones previas

- ¿Cuáles son los parámetros de interés en este modelo? y ¿Cómo modelar la incertidumbre sobre su localización?
 - $\beta_p \in \mathbb{R} \rightarrow \beta_p \sim \text{Normal}$
 - $\sigma_u^2 \in \mathbb{R}^+ \rightarrow \sigma_u^2 \sim \text{Gamma - inversa}$



Modelamiento bayesiano

Recordemos el modelo

$$\begin{aligned}\hat{\theta}_d^{\text{Dir}} &= \theta_d + \varepsilon_d & \varepsilon_d &\sim N(0, \text{Var}(\hat{\theta}_d^{\text{Dir}})) \\ \theta_d &= \mathbf{x}'_d \boldsymbol{\beta} + u_d & u_d &\sim N(0, \sigma_u^2)\end{aligned}$$

Verosimilitud

$$\hat{\theta}_d^{\text{Dir}} | \theta = \hat{\theta}_d^{\text{Dir}} | \boldsymbol{\beta}, \sigma_u^2 \sim N\left(\mathbf{x}'_d \boldsymbol{\beta} + u_d, \text{Var}(\hat{\theta}_d^{\text{Dir}}) + \sigma_u^2\right)$$

Distribuciones previas

$$\begin{aligned}\beta_1, \dots, \beta_p &\sim \text{Normal}(b_p, B_p) \\ \sigma_u^2 &\sim \text{IG}(\alpha, \delta)\end{aligned}$$

Distribución posterior

$$\theta_d | \hat{\theta}_d^{\text{Dir}} \sim N\left(\mathbf{x}'_d \boldsymbol{\beta} + u_d, \text{Var}(\hat{\theta}_d^{\text{Dir}}) + \sigma_u^2\right) \times N(b_1, B_1) \times \dots \times N(b_p, B_p) \times \text{IG}(\alpha, \delta)$$

Distribuciones atómicas

Nos encontramos frente a un cuello de botella, pues en general, las distribuciones posteriores bayesianas, no tienen una forma cerrada, ni conocida



El proyecto Manhattan

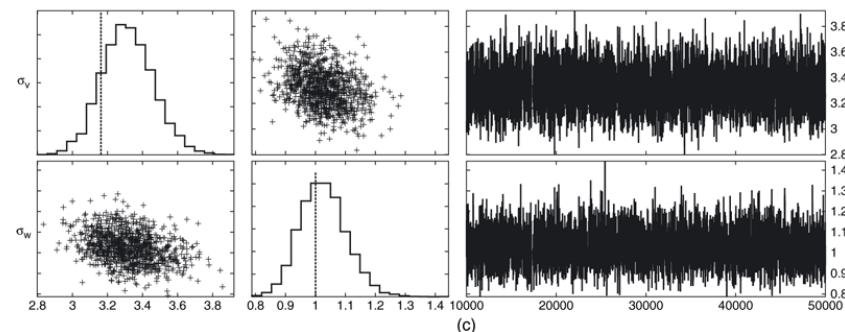
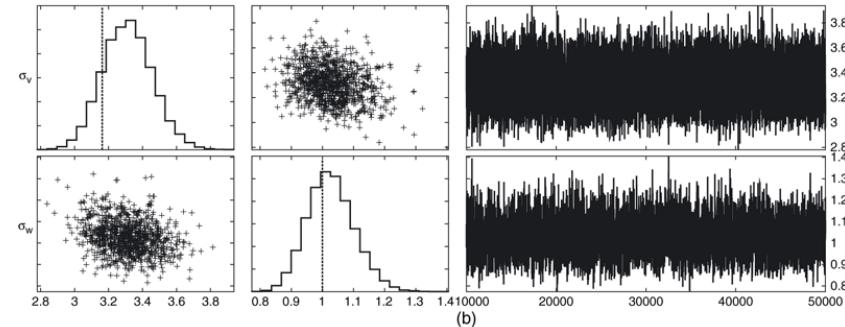
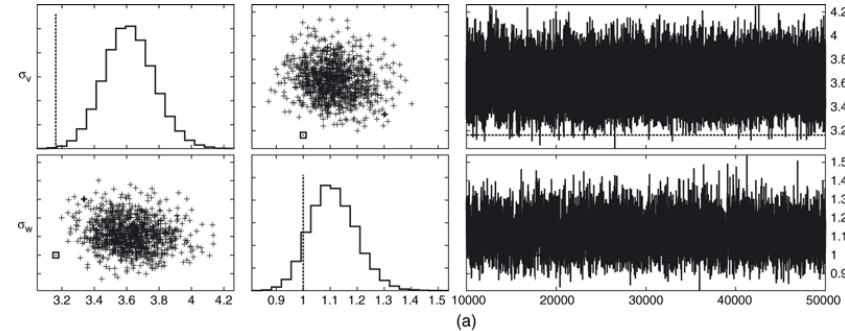
Paradójicamente, gracias al desarrollo de la bomba atómica se pudo dar solución al problema de las distribuciones difíciles:

- Stanislaw Marcin Ulam, pasó varios días jugando al solitario debido a una enfermedad.
 - Observó que era mucho más simple tener una idea del resultado general del juego haciendo pruebas múltiples y contando las proporciones de los resultados, que calcular todas las posibilidades.
- Esta idea la consiguió extrapolar a su trabajo en el Proyecto Manhattan en el que estaba teniendo problemas para calcular las ecuaciones integro-diferenciales que regían la difusión de neutrones.
 - El Proyecto Manhattan se desarrolló durante la Segunda Guerra Mundial para el Gobierno de los Estados Unidos, con la colaboración de Reino Unido y de Canadá.
 - El objetivo era desarrollar la primera bomba atómica
- A esta idea se le sumaron después otros científicos como John von Neumann y Metrópolis.

Los métodos MCMC

Al combinar un proceso estocástico como las cadenas de Markov con el método de Monte Carlo, se crea un método que es muy usado en la generación de números aleatorios.

- Se usa la simulación de Monte Carlo para crear una cadena de Markov con distribución estacionaria idéntica a la que se quiere conseguir.
 - Generación: se propone un punto candidato desde una distribución manejable y conocida.
 - Decisión: se acepta o se rechaza este punto candidato con una determinada probabilidad.
- Se repite este algoritmo hasta que la cadena converga en distribución.



JAGS

```
FH.model <- function(){
# Verosmilitud f(dir/theta)
for(j in 1:m1){
  thetahat[j] ~ dnorm(theta[j],
                        1/vhat.dir[j])
  theta[j] ~ dnorm(inprod(Beta[], X[j,]),
                    sigma2u.inv)
}
# Distribuciones previas f(theta)
sigma2u ~ dunif(0, 10^20)
sigma2u.inv <- pow(sigma2u, -1)
for (k in 1:p){
  Beta[k]~dnorm(0,0.0000000001)
}
# Distribución posterior predictiva
for(k in 1:m2){
  thetapred[k] ~ dnorm(inprod(Beta[], Xs[k,])
                        sigma2u.inv)
}
}
```

STAN

```
parameters {
  vector[p] beta;
  real<lower=0> sigma2_v;
  vector[N1] v;
}

transformed parameters{
  vector[N1] theta;
  real<lower=0> sigma_v;
  theta = X * beta + v;
  sigma_v = sqrt(sigma2_v);
}

model {
  beta ~ normal(0, 100);
  sigma2_v ~ inv_gamma(0.0001, 0.0001);
  y ~ normal(theta, sigma_e);
  v ~ normal(0, sigma_v);
}

generated quantities{
  vector[N2] y_pred;
  for(j in 1:N2) {
    y_p[j] = normal_rng(Xs[j] * beta,
                          sigma_v);}
}
```

Inferencia posterior

Inference for Stan model: FH_normal.

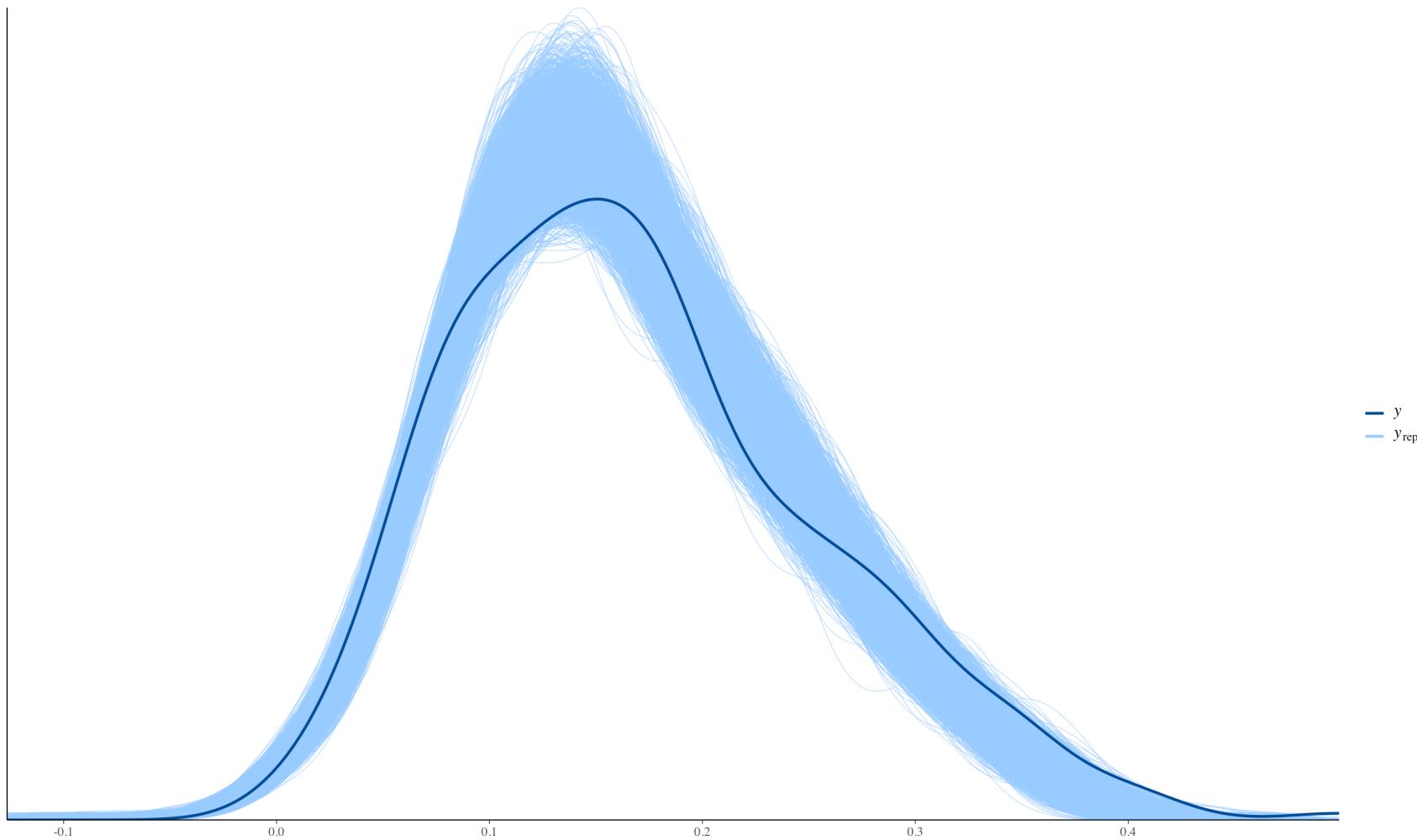
4 chains, each with iter=2000; warmup=1000; thin=1;

post-warmup draws per chain=1000, total post-warmup draws=4000.

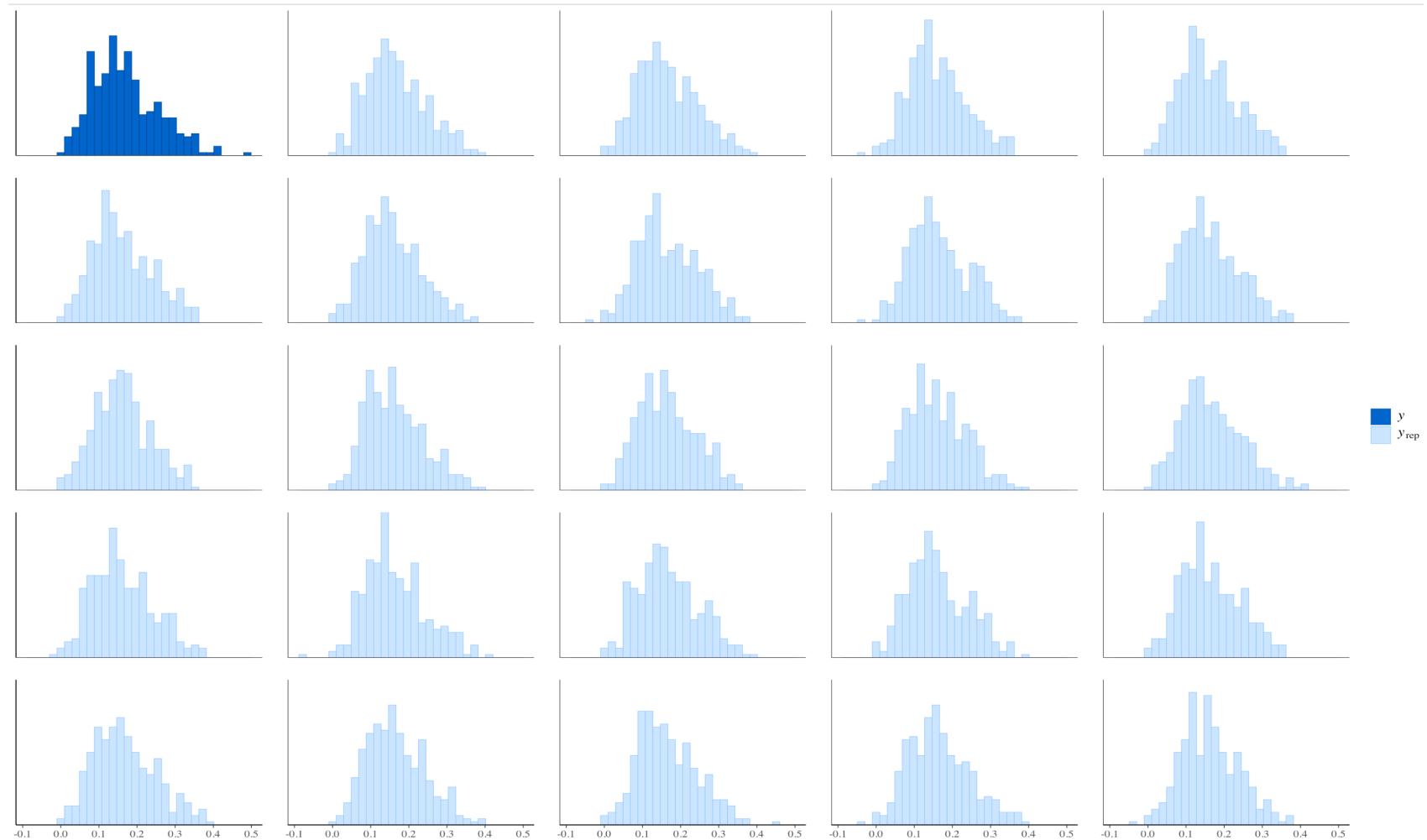
	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%
sigma2_v	0.0015	0.0000	0.0002	0.0012	0.0014	0.0015	0.0016	0.0019
beta[1]	-0.0920	0.0016	0.0189	-0.1269	-0.1057	-0.0928	-0.0788	-0.0538
beta[2]	-0.1163	0.0007	0.0162	-0.1482	-0.1269	-0.1167	-0.1055	-0.0844
beta[3]	-0.2464	0.0049	0.1196	-0.4747	-0.3283	-0.2446	-0.1658	-0.0127
beta[4]	0.6720	0.0063	0.1624	0.3524	0.5611	0.6755	0.7801	0.9967
beta[5]	-0.2382	0.0062	0.1445	-0.5224	-0.3344	-0.2383	-0.1430	0.0475
beta[6]	-0.0184	0.0083	0.1711	-0.3649	-0.1310	-0.0162	0.0962	0.3126
beta[7]	1.5730	0.0234	0.2783	1.0222	1.3773	1.5885	1.7764	2.0806
beta[8]	1.5388	0.0235	0.2791	0.9848	1.3447	1.5555	1.7396	2.0465
beta[9]	1.5627	0.0234	0.2790	1.0154	1.3674	1.5787	1.7648	2.0682
beta[10]	1.5841	0.0231	0.2771	1.0335	1.3904	1.5988	1.7850	2.0870

Convergencia de las cadenas

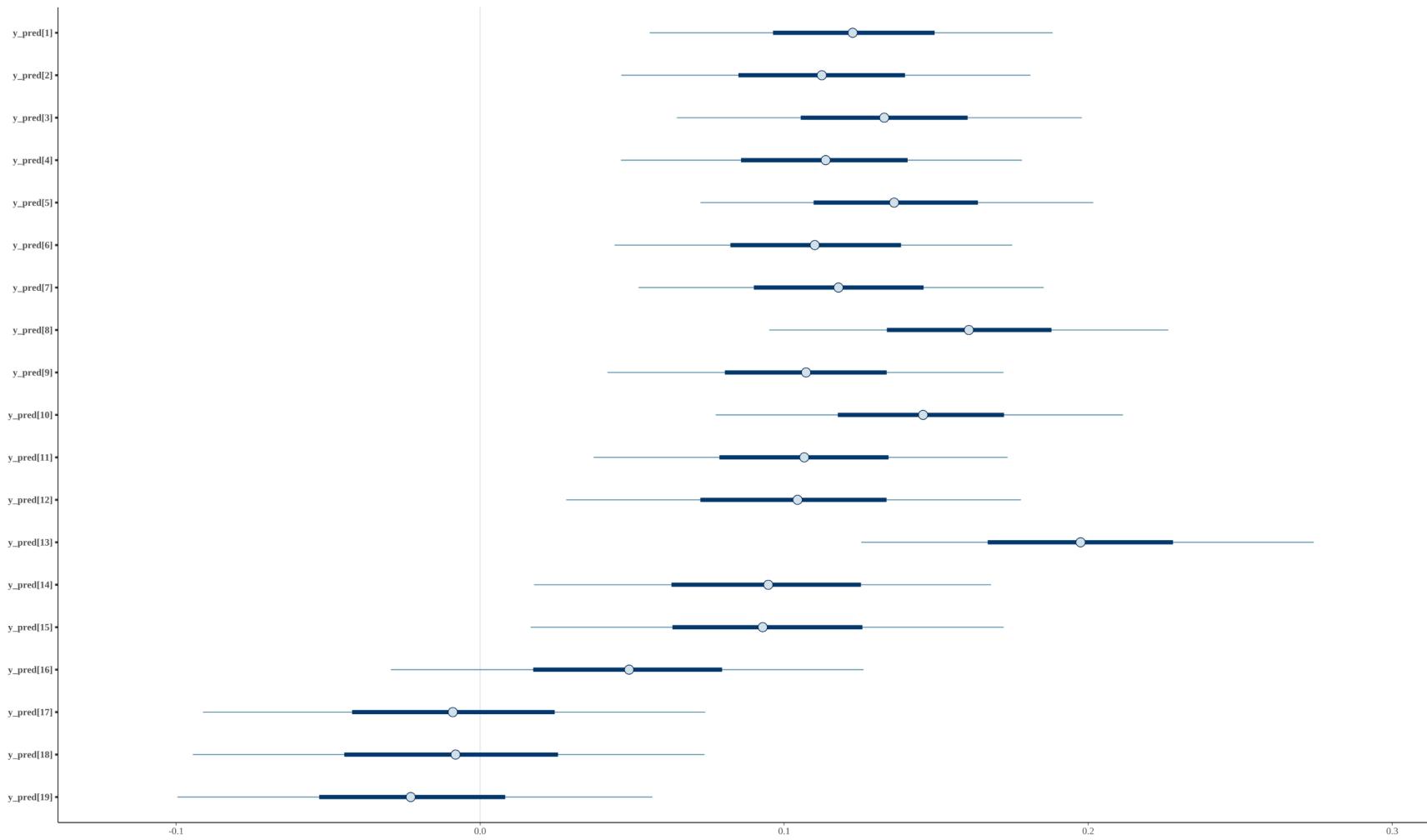
Calidad del modelo



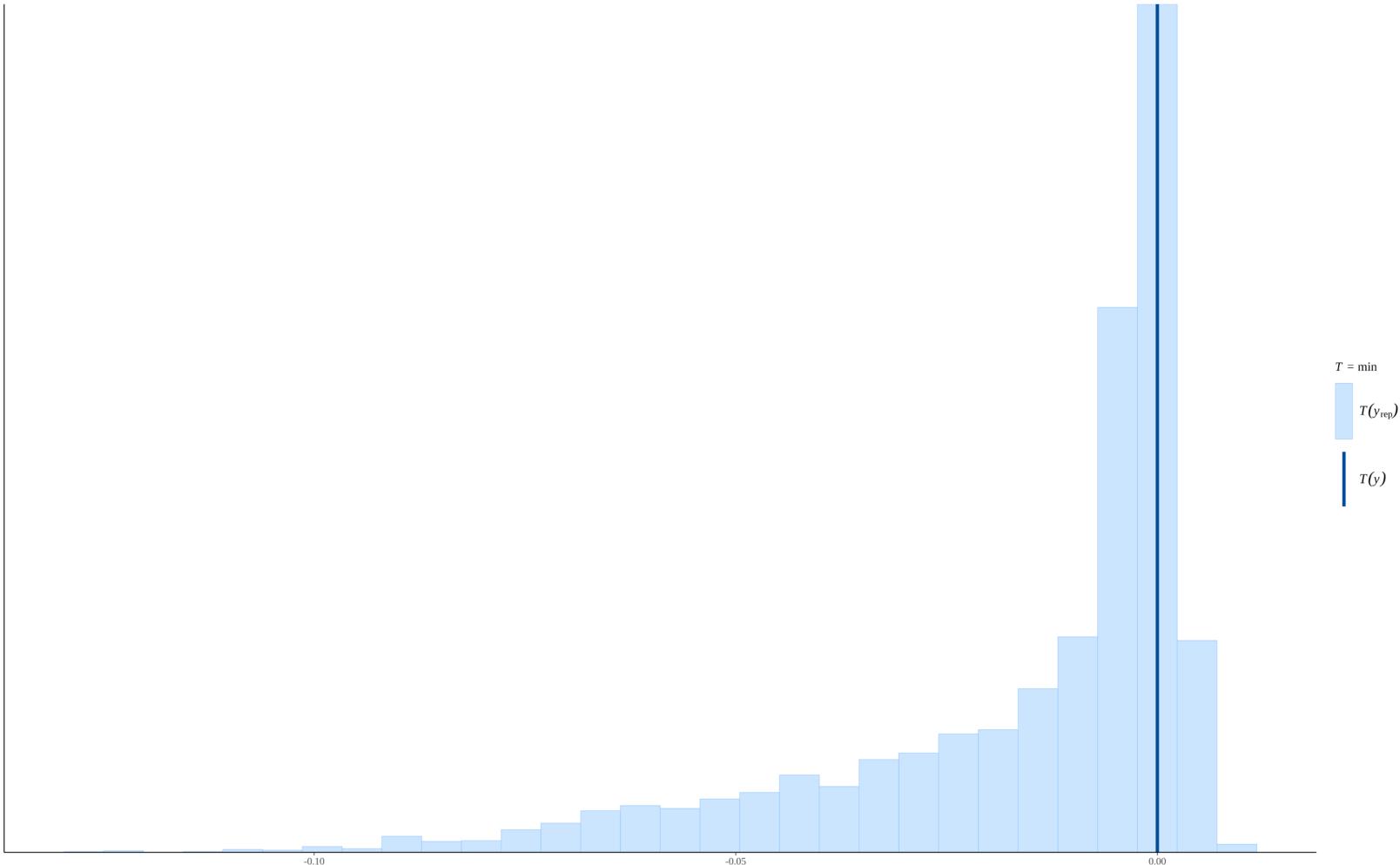
Calidad del modelo



Calidad predictiva



Expectativa vs. realidad



Un mejor modelo para las proporciones

La distribución beta

Si el parámetro de interés es una proporción $\in (0, 1)$, ¿por qué proponer una distribución previa normal $\in (-\infty, \infty)$?

Un mejor modelo para las proporciones

Model 4: (beta-logistic model with unknown sampling variance)

Sampling model:

$$p_{iw} | P_i \stackrel{ind}{\sim} beta(a_i, b_i) \quad (3.8)$$

Linking model:

$$\text{logit}(P_i) | \beta, \sigma_v^2 \stackrel{ind}{\sim} N(x_i' \beta, \sigma_v^2) \quad (3.9)$$

For both Model 3 and Model 4, the approximate variance function $\psi_i = [P_i(1 - P_i)/n_i]deff_{iw}$ is used. The parameters a_i and b_i in Model 4 are given by:

$$a_i = P_i \left(\frac{n_i}{deff_{iw}} - 1 \right), \text{ and } b_i = (1 - P_i) \left(\frac{n_i}{deff_{iw}} - 1 \right).$$

Modelo 4 - (Liu, Lahiri, Kalton, 2014)

Verosimilitud

$$\begin{aligned}\hat{\theta}_d^{\text{Dir}} | \theta &\sim \text{Beta}(a_d, b_d) \\ \text{logit}(\theta) | \beta, \sigma_u^2 &\sim \text{Normal}(\mathbf{x}'_d \beta, \sigma_u^2)\end{aligned}$$

$$a_d = \theta_d(n_{\text{eff}} - 1) \quad , \quad b_d = (1 - \theta_d) \times (n_{\text{eff}} - 1)$$

Distribuciones previas

$$\begin{aligned}\beta_1, \dots, \beta_p &\sim \text{Normal}(b_p, B_p) \\ \sigma_u^2 &\sim \text{IG}(\alpha, \delta)\end{aligned}$$

Distribución posterior

$$\begin{aligned}\theta_d | \theta_d^{\text{Dir}} &\sim \text{Beta}(a_d, b_d) \times \text{NormalLogitInv} \left(\mathbf{x}'_d \beta + u_d, \text{Var}(\hat{\theta}_d^{\text{Dir}}) + \sigma_u^2 \right) \\ &\times N(b_1, B_1) \times \cdots \times N(b_p, B_p) \times \text{IG}(\alpha, \delta)\end{aligned}$$

En STAN

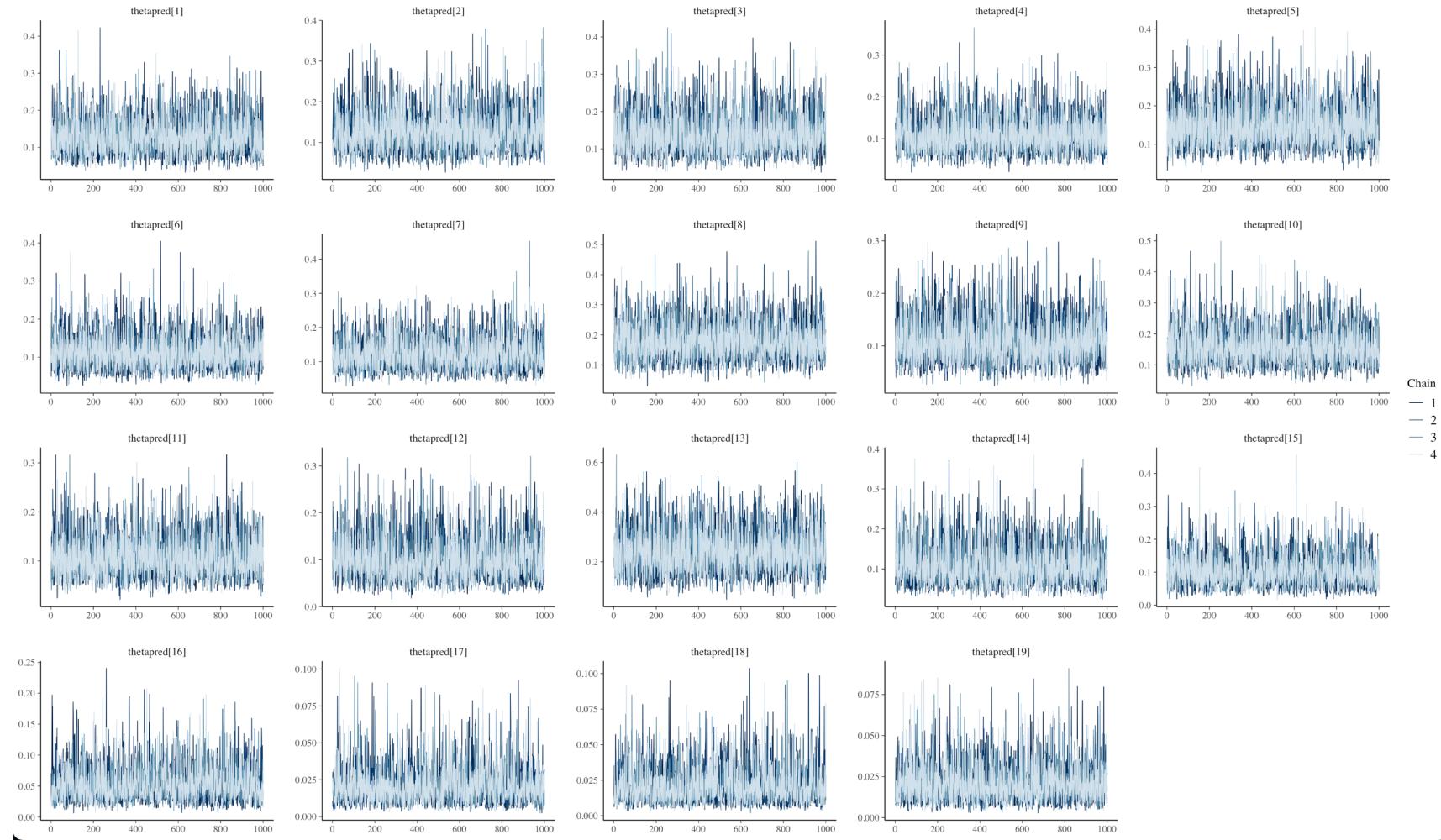
```
parameters {
  vector[p] beta;
  real<lower=0> sigma2_v;
  vector[N1] v;
}

transformed parameters{
  vector[N1] LP;
  real<lower=0> sigma_v;
  vector[N1] theta;
  LP = X * beta + v;
  sigma_v = sqrt(sigma2_v);
  for (i in 1:N1) {
    theta[i] = inv_logit(LP[i]);
  }
}
```

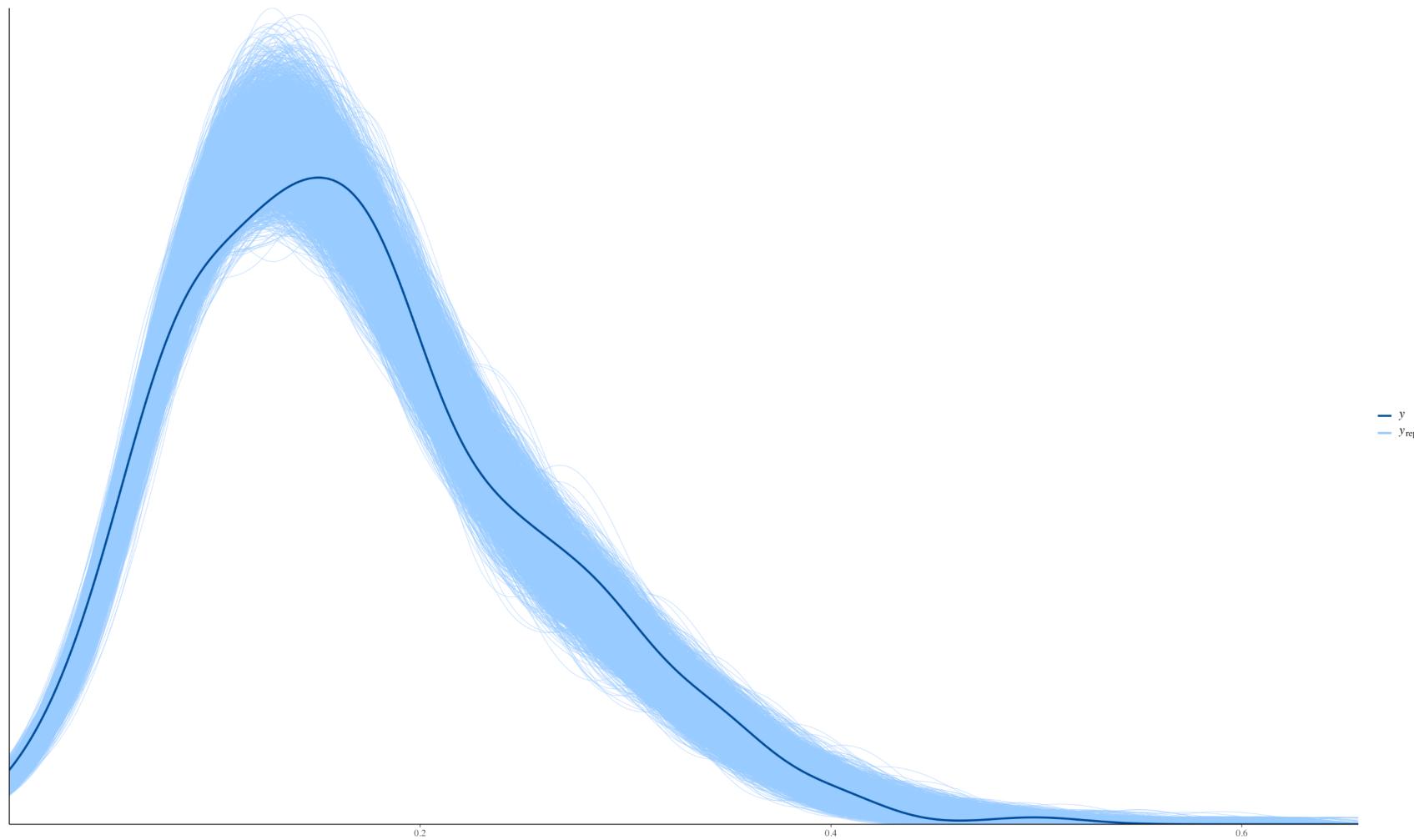
```
model {
  vector[N1] a;
  vector[N1] b;
  for (i in 1:N1) {
    a[i] = theta[i] * phi[i];
    b[i] = (1 - theta[i]) * phi[i];
  }
  // priors
  beta ~ normal(0, 100);
  sigma2_v ~ inv_gamma(0.0001, 0.0001);
  // likelihood
  y ~ beta(a, b);
  v ~ normal(0, sigma_v);
}

generated quantities {
  vector[N2] y_pred;
  vector[N2] thetapred;
  for (i in 1:N2) {
    y_pred[i] = normal_rng(Xs[i] * beta,
                           sigma_v);
    thetapred[i] = inv_logit(y_pred[i]);
  }
}
```

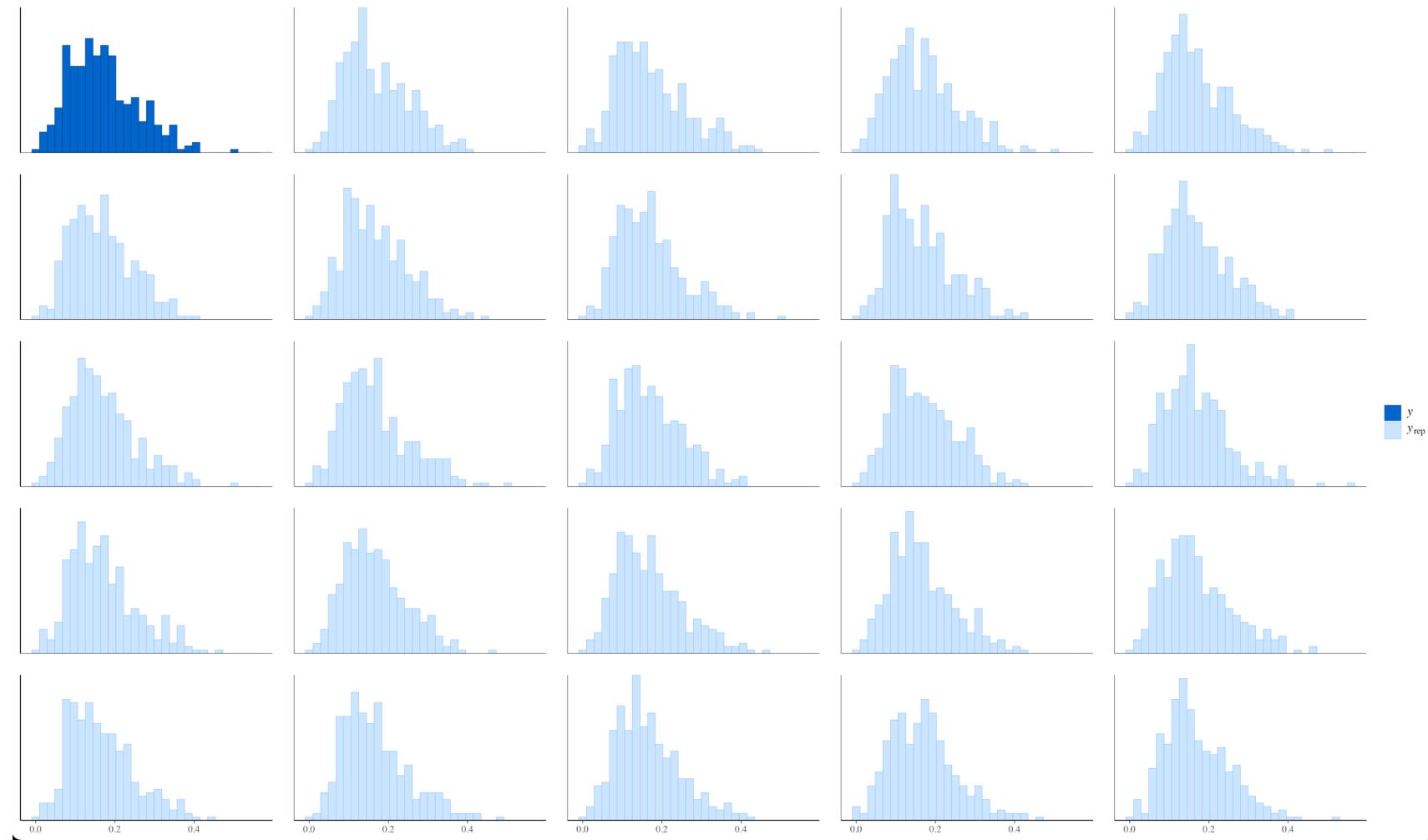
Cadenas de la predicción de comunas fuera de la muestra



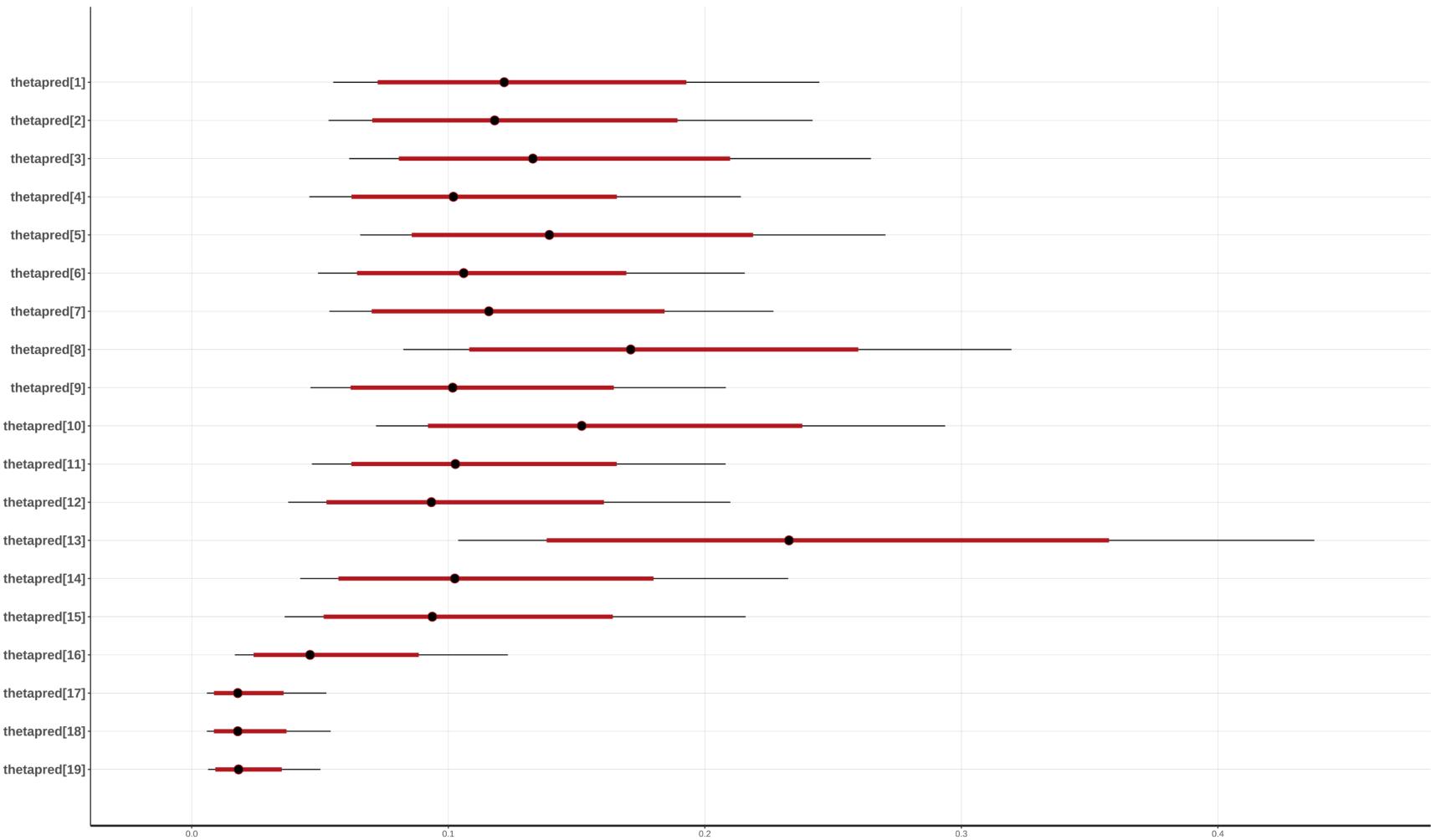
Calidad del modelo



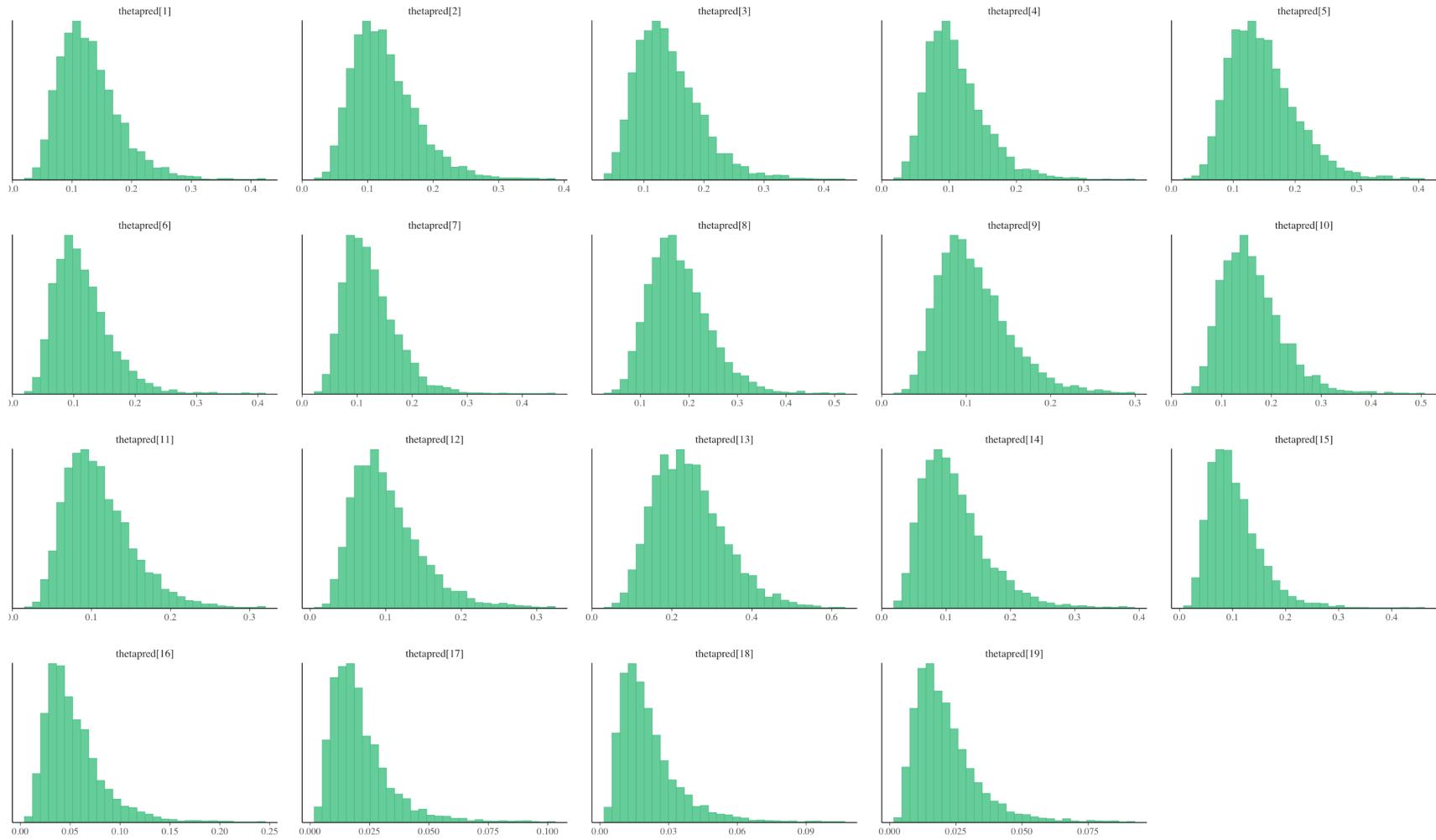
Calidad del modelo



Calidad predictiva



Distribución posterior predictiva



Démosle un vistazo I

Démosle un vistazo II

¡Gracias!

andres.gutierrez@un.org