QÜESTIIÓ, vol. 21, 1 i 2, p. 59-97, 1997

DISEÑO Y ANÁLISIS BAYESIANOS DE RÉPLICAS EN LA EXPERIMENTACIÓN CIENTÍFICA*

M.J. BAYARRI*
M.A. MARTÍNEZ*
Universitat de València

En la práctica estadística es frecuente la replicación de experiencias estadísticas, justificada por muy diversos intereses: ratificar conclusiones, estudiar la variación en las respuestas cuando se experimenta sobre una población distinta, validar un modelo, detectar sesgos, etc. Al disponer, antes de replicar, de los resultados de cierto estudio previo, es razonable pensar en buscar un buen diseño para la réplica utilizando la información que dicho primer estudio proporciona. Un diseño óptimo de la réplica consistirá en decidir, asumido un modelo, sobre el menor tamaño muestral que otorga al investigador suficientes garantías de concluir con éxito en su inferencia. Una modelización jerárquica bayesiana permite precisar fácilmente las relaciones entre datos y poblaciones (original y replicada) y derivar conclusiones sobre los resultados más probables (en función del tamaño de la réplica) a través de las distribuciones a posteriori y predictivas que se obtengan.

Bayesian design and analysis of replications in scientific experimentation

Keywords: Replicación, modelo jerárquico bayesiano, factor Bayes, simulación, aproximación Monte-Carlo.

⋆Este trabajo ha sido parcialmente subvencionado por la Consellería de Cultura, Educació i Ciència de la Generalitat Valenciana, bajo el proyecto de Investigación GV-1081/93.

*M.J. Bayarri y M.A. Martínez. Dpto. Estadística e I.O. Universitat de València. Dr. Moliner, 50. 46100 BURJASSOT. València.

– Article rebut el desembre de 1995.

– Acceptat el juny de 1996.

1. INTRODUCCIÓN

1.1 La replicación de experimentos. Contexto

A pesar de que la ***replicación*** es una práctica estadística habitual en la investigación experimental, no parece existir una modelización sistemática encaminada a sacar el máximo provecho de la información que puede proporcionar; en la bibliografía sólo hemos encontrado algunos apuntes acerca de su potencial en la investigación y diversas consideraciones sobre la interpretación «acertada» de los resultados de la réplica (ver Neuliep, 1990).

Entendemos por ***réplica*** o ***replicación*** la repetición de una experiencia estadística ε_o, consistente en una nueva recogida de datos y análisis de los mismos, con el objetivo último de inferir sobre cierto parámetro poblacional (referente a una media, una diferencia de medias, de efectos de tratamientos, un coeficiente de correlación, etc).

Para replicar una experiencia estadística previa, el investigador elegirá, en función del objetivo de su estudio, la misma población muestreada en ε_o u otra diferente; decidirá asimismo entre la reproducción de las condiciones de experimentación originales (replicación exacta) y su modificación (parcial o total), para inferir sobre la variabilidad de la característica que estudia.

Se hace uso de la replicación en todos aquellos campos de investigación en los que es común la repetición de análisis estadísticos para profundizar en el estudio de cierta característica poblacional; campos tan diversos como Medicina, Ecología, Economía, Sociología, Agricultura, Mejora Animal, etc.

1.2 Interpretación

A pesar de lo ampliamente debatido que ha sido el tema de la replicación (ver Neuliep, 1990), no existe un criterio claro sobre cómo valorar los resultados que las réplicas proporcionan. Para ilustrar esta problemática Tversky y Kahnemann (ver Utts, 1991) planteaban la replicación de una experiencia basada en la resolución de un contraste *t* sobre cierto parámetro de interés θ (contraste que nosotros adaptamos, por simplificar, a uno basado en una población *normal* con media θ y varianza conocida e igual a 1):

$$\begin{cases} H_0 & : \quad \theta \leq 0 \\ H_1 & : \quad \theta > 0. \end{cases}$$

A partir de una primera muestra de 15 observaciones, se había obtenido un valor para el estadístico de contraste de:

$$z_o = 2.19.$$

(La media muestral observada había sido $x_o = 0.5654$).

A continuación se formulaba la siguiente pregunta a un grupo de expertos: «Si replicáramos con el mismo tamaño muestral ($n_r = 15$), ¿cuál sería el máximo valor del estadístico z_r que podría ser considerado como ***fracaso*** al reproducir el primer estudio?».

La respuesta general fue un valor de:

$$z_r = 1.60.$$

Como hacían notar Tversky y Kahnemann (a los que nos referiremos por T&K a partir de ahora), las implicaciones de este estudio empírico resultaban un poco contradictorias. Por una parte, mientras que el resultado de la réplica ($z_r = 1.60$) se percibía a primera vista como un fracaso a la hora de intentar reproducir el experimento original, reduciendo así nuestra confianza en las conclusiones del primer estudio, si decidiéramos combinar los datos del primer y segundo estudios como pertenecientes a una misma muestra (llevando a cabo un análisis global sobre el efecto en cuestión), conseguiríamos sin embargo, ratificar con más fuerza las conclusiones de dicho primer estudio, al obtener para el estadístico combinado z un valor superior al de z_o:

$$z = 2.68,$$

en contra entonces, de lo que en principio hacía suponer la conclusión individual de la réplica.

La forma más adecuada de concluir sobre lo observado en la réplica y cómo definir entonces la consecución de «éxito» o fracaso al replicar, dependerá de cada problema específico. En ocasiones interesará combinar toda la información disponible (estudio original y réplica), mientras que en otras el interés residirá en la comparación de las conclusiones obtenidas individualmente en uno y otro estudios.

Para abordar la sistematización de las réplicas asumimos que el investigador ya ha planteado:

(a) cuál es el problema concreto que pretende resolver con la réplica, esto es, cuáles son los ***objetivos*** inmediatos con que justifica esta «repetición» estadística;

(b) cómo va a replicar: qué observará, qué modelo propondrá y al fin, qué metodología y procedimiento estadísticos seguirá para concluir en su estudio.

Los ***objetivos*** del experimentador para emprender la replicación de un experimento pueden ser muy diversos. Basados en las referencias mencionadas sobre replicación y en algunas indicaciones encontradas en réplicas publicadas, hemos intentado caracterizar los más frecuentes e importantes (ver Mayoral, 1995). Entre ellos podríamos resaltar:

1. la ***confirmación*** de las conclusiones obtenidas en el primer estudio, esto es, la validación (ratificación) de su inferencia, replicando sobre la misma población de ε_o y bajo idénticas condiciones de experimentación;
2. la ***generalización*** de dichas conclusiones a otras poblaciones distintas a la muestreada en ε_o;
3. el estudio de la ***variabilidad*** de la característica de interés en función de la evolución de la población y/o las modificaciones experimentales introducidas respecto de ε_o; en definitiva, lograr concluir sobre el dominio de las conclusiones primeras y su sensibilidad a los cambios en la experimentación;
4. la ***distinción*** de poblaciones y/o ***discriminación*** de variables en la explicación de cierto efecto poblacional de interés, detectable únicamente cuando se repiten experimentos modificando de alguna forma las variables que los afectan, o simplemente alterando su valor (replicaciones no exactas);
5. la ***detección de sesgo*** en un primer estudio, ocasionado por la ausencia o deficiencias del diseño.

Aunque no es crucial al desarrollo de nuestro trabajo, supondremos por simplicidad que la réplica aborda un problema estadístico similar al del estudio original (fuera éste un contraste de hipótesis, un problema de estimación, etc), utilizando para su resolución la misma metodología (clásica, bayesiana, ...) y procedimiento (p-valores, estimadores, distribuciones a posteriori, probabilidades finales, factores Bayes, ...) empleados en el primer estudio.

La estadística bayesiana hace muy fácil la incorporación secuencial de información, hecho valorable en todo problema que involucra replicaciones sucesivas de un primer experimento de las que ir aprendiendo. Asimismo, permite una modelización sencilla para las relaciones entre los distintos experimentos por medio de las distribuciones de los parámetros (modelo jerárquico). La inferencia y la predicción se obtienen de una manera natural a partir de las distribuciones a posteriori y predictivas respectivamente. Constituye así la replicación, un marco idóneo para la aplicación de la metodología bayesiana. La consideración de los objetivos, inferencias y características particulares de cada investigación conducirán nuestra modelización y análisis bayesianos de las réplicas.

2. PLANTEAMIENTO DEL PROBLEMA

Llamaremos ε_r a la replicación de cierto análisis estadístico previo ε_o. X_o y X_r representarán las variables a observar en cada uno de ellos (en nuestro caso, estadísticos suficientes; en general identificarán a vectores de observaciones).

Asumimos el estudio sobre cierto parámetro poblacional, conceptualmente similar en ambos experimentos (una media para nosotros), representado por θ_o en el estudio original y θ_r en la réplica.

Trabajamos únicamente bajo tres supuestos básicos (una relación más exhaustiva junto con su correspondiente análisis estadístico puede encontrarse en Mayoral, 1995).

S1. ε_o consiste en un contraste de hipótesis sobre el parámetro poblacional θ_o. Se propone el mismo contraste en la réplica, sobre una población *similar* a la de ε_o, con el objetivo de ***reproducir*** la conclusión original acerca del rechazo (o no) de la hipótesis nula.

S2. En ε_o se infiere sobre un parámetro de interés θ_o. La réplica, llevada a cabo generalmente sobre una población distinta, se plantea para inferir sobre la ***discrepancia*** entre ambas poblaciones, esto es, sobre la diferencia de efectos $\delta = \theta_r - \theta_o$.

S3. Se sospecha la ***existencia de sesgo*** (β) en un primer estudio en el que se pretendía inferir sobre cierto parámetro μ. Suponiendo poder controlar dicho sesgo en un estudio posterior sobre la misma población, se trata de inferir sobre él.

La flexibilidad del ***modelo jerárquico bayesiano*** permite una fácil modelización de las relaciones que surgen en replicación: la discrepancia entre los parámetros θ_o y θ_r ocasionada por el distanciamiento temporal, espacial, poblacional, etc, entre el primer estudio y la réplica; la presencia de ruidos no evitables en la implementación y que afectan a la percepción del efecto real a través de la muestra; la variación del modelo que las modificaciones experimentales en la réplica pueden ocasionar (inclusión/exclusión de variables de influencia, alteración del rango de valores posibles para las variables y parámetros involucrados,...), etc.

Utilizamos un modelo jerárquico sencillo con únicamente tres niveles: en el primer nivel modelizamos la distribución condicional de las variables observables (X_o y X_r); en el segundo nivel, las relaciones entre los parámetros que definen al modelo en el primer nivel (θ_o y θ_r); el tercer nivel es reservado para especificar las distribuciones a priori de los parámetros restantes (hiper-parámetros).

Trabajando con metodología bayesiana derivamos las correspondientes ***distribuciones a posteriori*** de interés, a través de las cuales podremos medir la similitud o

discrepancia entre los estudios original y réplica, y una ***distribución predictiva*** para el resultado de la réplica que nos permitirá, entre otras cosas, la planificación «óptima» de ésta.

En función del objetivo del investigador y de la inferencia a realizar se proponen distintas posibilidades de definir el ***éxito al replicar***. Cada definición de éxito se traduce en cierta condición que habrá de satisfacer el resultado de la réplica, x_r. Puesto que éste aún no se ha observado (todo el análisis es previo a la replicación), la probabilidad de éxito se calculará utilizando la distribución de la variable aleatoria X_r. Dada la información proporcionada por el primer estudio, esta distribución es la predictiva $m(x_r|x_o)$, dependiente del tamaño muestral de la réplica, n_r. El análisis derivado del modelo propuesto proporcionará herramientas para diseñar una réplica (decidir un tamaño n_r) que otorgue suficientes «garantías de éxito», o simplemente sugerirá desestimar su práctica en caso de no rebasar éstas el nivel exigido. Fijado por el investigador dicho nivel de garantías M, el tamaño «óptimo» para replicar se obtendrá para el menor valor de n_r que proporcione una probabilidad de éxito tal que:

$$Prob(\text{ éxito }) \geq M. \tag{1}$$

Estudiamos cada uno de los supuestos (S1, S2 y S3) modelizando y obteniendo las inferencias necesarias para, una vez planteado en cada caso lo que se entenderá por éxito, calcular la probabilidad de lograrlo con los datos del ejemplo propuesto por T&K sobre una «posible» replicación de un problema de contraste de hipótesis (planteado al inicio de la sección 1.2).

Introduciremos dos modelos que denotaremos por M12 y M3. El modelo M12 será la base para estudiar la reproducción de conclusiones en el contraste (S1) y para comparar las poblaciones (S2). El modelo M3 nos permitirá inferir sobre el sesgo «intuído» en el primer estudio (S3). Para ambos modelos se incorporará en la última parte de este trabajo, un grado más de incertidumbre (modelos M12A y M3A), con el objeto de robustecer los resultados ya obtenidos tras flexibilizar la modelización.

3. REPRODUCCIÓN Y COMPARACIÓN CON LA RÉPLICA

3.1. Modelo M12

Planteamos la repetición de cierta experiencia estadística previa, para lo cual asumimos la modelización siguiente:

I. DATOS

Suponemos que las muestras en el estudio original y en la réplica son (condicionalmente) independientes, y que en cada estudio el estadístico suficiente X_i tiene una distribución normal con varianza conocida:

$$\begin{aligned} f(x_o, x_r|\theta_o, \theta_r) &= f(x_o|\theta_o) \cdot f(x_r|\theta_r) \\ f(x_i|\theta_i) &= N(x_i|\theta_i, \gamma_i^2), \quad i = o, r, \end{aligned} \tag{2}$$

con $\gamma_i^2 = \sigma^2/n_i$, σ^2 conocida (=1 por simplificar) y n_i ($i = o, r$) el tamaño muestral utilizado en cada estudio.

II. PARÁMETROS

Suponemos (como una primera aproximación sencilla a la posible relación entre los dos estudios) que los parámetros poblacionales en el experimento original y en la réplica son intercambiables (la consideración de relaciones más complejas entre ambos experimentos —involucrando covariables, relaciones espaciales y/o temporales, etc—, se abordará en trabajos futuros):

$$\begin{aligned} p(\theta_o, \theta_r|\mu) &= p(\theta_o|\mu) \cdot p(\theta_r|\mu) \\ p(\theta_i|\mu) &= N(\theta_i|\mu, \tau^2), \quad i = o, r. \end{aligned} \tag{3}$$

La varianza τ^2, que suponemos conocida, cuantifica la información a priori acerca de la semejanza o disparidad entre los dos estudios: cuanto menor es τ^2, más homogéneas se consideran las poblaciones del experimento original y de la réplica. Estudiaremos en nuestro ejemplo de T&K la sensibilidad de las conclusiones a la magnitud de τ^2 utilizando un valor moderadamente alto, $\tau^2 = 1$, quince veces mayor que la varianza muestral del estudio original, y otro moderadamente pequeño, $\tau^2 = 1/5$, tan sólo tres veces mayor que $\gamma_o^2 (= 1/15)$.

III. HIPER-PARÁMETRO

Utilizamos una distribución mínimo-informativa a priori para el hiper-efecto poblacional μ (que puede interpretarse, si se desea, como la media de una hiperpoblación de la que «muestreamos» θ_o y θ_r):

$$p(\mu) \propto \text{ cte} . \tag{4}$$

A la modelización clásica (primer nivel) se añade el elemento bayesiano $p(\theta_o, \theta_r)$ que modeliza la relación entre los dos experimentos. Tendremos así acceso tanto a las

conclusiones derivadas de análisis clásicos en ε_o y ε_r, como a las obtenidas de análisis bayesianos. Obviamente, resulta más natural llevar a cabo un análisis coherente y concluir sobre los parámetros de ambos estudios en base a sus correspondientes distribuciones a posteriori. Sin embargo, el análisis que proponemos es más amplio y podrá aplicarse también cuando las inferencias que se deseen realizar acerca de θ_r estén basadas en la metodología clásica. El hecho clave para inferir sobre la réplica es que su resultado X_r es una variable aleatoria a la que el modelo anterior permite dotar de una distribución predictiva, obtenida de las distribuciones finales para los parámetros:

$$\begin{aligned} m(x_r|x_o) &= \int f(x_r|\theta_r)\cdot p(\theta_r|x_o)\, d\theta_r \\ &= N(x_r|x_o, \gamma_o^2+\gamma_r^2+2\tau^2), \end{aligned} \tag{5}$$

donde $p(\theta_r|x_o)$ contiene toda la información sobre el parámetro poblacional θ_r proporcionada por el primer estudio:

$$p(\theta_r|x_o) = \int p(\theta_r|\mu)\cdot p(\mu|x_o)\, d\mu,$$

con

$$p(\mu|x_o) \propto f(x_o|\mu)\cdot p(\mu) \propto \int f(x_o|\theta_o)\cdot p(\theta_o|\mu)\, d\theta_o.$$

3.2. Reproducción de conclusiones

Suponemos realizado un primer análisis estadístico ε_o que resolvía cierto contraste de hipótesis sobre una media poblacional. Se propone un segundo muestreo ε_r sobre una población que a priori se supone «próxima» a la inicial (podría tratarse de la misma) y, manteniendo fijas las condiciones de experimentación originales, se pretende reproducir las primeras conclusiones. Nótese que cuando la replicación se llevara a cabo sobre la misma población de ε_o, lograr este objetivo significaría «validar» el primer análisis. La desviación aleatoria entre las medias poblacionales θ_o y θ_r viene descrita por el modelo de intercambiabilidad. El contraste que en ε_r se plantea será análogo al del estudio original:

$$\begin{cases} H_0^i & : \quad \theta_i \le c_i \\ H_1^i & : \quad \theta_i > c_i, \quad i = o, r. \end{cases}$$

Consideramos $c_r = c_o = c$, esto es, pretendemos concluir de la misma forma sobre la magnitud de θ_r que sobre la de θ_o. Sin pérdida de generalidad tomamos $c = 0$, puesto que cualquier contraste de este tipo lo podemos desplazar en torno al cero sin más que restar una constante.

En el ejemplo de T&K, la inferencia en el primer estudio se llevaba a cabo a través del p-valor observado (metodología clásica). Al observar en él un valor para

el estadístico de contraste de $z_o = 2.19$, se rechazaba la hipótesis nula a un nivel de significatividad $\alpha = 0.05$. Reproducir conclusiones significará entonces, ***volver a rechazar*** la hipótesis nula en la réplica:

$$\begin{aligned} \textbf{ÉXITO} &\equiv \textbf{Reproducir } \varepsilon_o \\ &\equiv \textbf{Rechazar } H_0^r. \end{aligned}$$

Una vez definido lo que consideraríamos ***éxito*** en la réplica, buscamos el «mejor» tamaño para lograrlo, a partir del análisis bayesiano derivado del modelo M12.

El rechazo de la hipótesis nula en la réplica admite sin embargo dos posibilidades (la disyuntiva que T&K cuestionaban):

(R1) Si se decide analizar la réplica de forma aislada, sin añadir la información que proporciona el primer estudio (primer supuesto en la situación paradójica de T&K), «***rechazar***» significará, siguiendo la metodología clásica para inferir en la réplica, obtener un p-valor p_r inferior a 0.05. En el escenario en que trabajamos, el p-valor (inferencia clásica) coincide con la probabilidad a posteriori de H_0^r (inferencia bayesiana). Así, el rechazo de H_0^r será la conclusión tanto de un clásico, como de un bayesiano que asumiera una función de pérdida que hiciera razonable rechazar dicha hipótesis cuando su probabilidad final resultara inferior a α:

$$p_r \leq \alpha \Leftrightarrow P(H_0^r|x_r) \leq \alpha.$$

Dada su equivalencia, es suficiente considerar uno de estos índices de evidencia para llevar a cabo nuestro análisis. Así, en la réplica tendremos ***éxito R1*** cuando:

$$p_r \leq 0.05. \tag{6}$$

Puesto que p_r depende del resultado de la réplica (también de su tamaño), $p_r = p_r(X_r, n_r)$, antes de replicar es una variable aleatoria cuya distribución viene determinada por la distribución de X_r. La distribución predictiva $m(x_r|x_o)$ será la herramienta natural para calcular la probabilidad de que la condición [6] se cumpla:

$$\begin{aligned} Prob(\text{ éxito R1 }) = P[p_r \leq \alpha|x_o] &= P[X_r \geq \frac{z_{1-\alpha}}{\sqrt{n_r}}|x_o] \\ &= \int_{z_{1-\alpha}/\sqrt{n_r}}^{+\infty} m(x_r|x_o)\, dx_r. \end{aligned}$$

(R2) Si se pretende combinar el resultado del estudio original con el de la réplica (segundo supuesto de T&K), es preciso especificar el tipo de análisis a realizar. Aunque obviamente es posible llevar a cabo un análisis clásico utilizando un p-valor combinado, nos decantamos por un análisis bayesiano que involucra toda la información disponible mediante las distribuciones a posteriori. Por mantener la analogía con R1, hablaremos de conseguir ***éxito R2*** rechazando H_0^r cuando la probabilidad final para la hipótesis nula una vez observada la réplica sea inferior a 0.05, esto es:

$$P(\theta_r \leq 0 | x_o, x_r) \leq 0.05. \tag{7}$$

De nuevo, fijado n_r obtendremos un rango de valores de X_r para los que se verifica la condición [7], y de ahí se podrá elegir el tamaño que otorgue suficientes «garantías de éxito» (una probabilidad de éxito razonablemente alta):

$$Prob(\text{ éxito R2 }) = \int_{\{x_r:\ P(\theta_r \leq 0 | x_o, x_r) \leq 0.05\}} m(x_r | x_o)\ dx_r\ \geq\ M,$$

con M el nivel de garantías exigido para replicar (impuesto por el investigador).

La elección de uno u otro tipo de rechazo en la inferencia a realizar con la réplica dependerá del investigador. Las probabilidades de éxito según R1 y R2 son bastante similares, como puede apreciarse en la Figura 1b, si bien el rechazo de H_0^r combinando información (R2) resulta algo más probable que el rechazo con sólo el resultado de la réplica (R1). Para una réplica de igual tamaño que ε_o ambas probabilidades de rechazo giran en torno a 0.5, sensiblemente inferiores a lo que intuitivamente cabía esperar.

Puesto que suponemos una ***replicación fiel*** del primer estudio, es razonable asumir a priori una ***discrepancia pequeña*** entre los parámetros poblacionales, es decir, un valor pequeño para $\tau^2 (= 1/5)$. Hemos hecho sin embargo, un estudio comparativo entre los resultados que obtendríamos partiendo de un valor pequeño ($\tau^2 = 1/5$) y uno grande ($\tau^2 = 1$), con el fin de estudiar la sensibilidad de las conclusiones respecto de la opinión inicial incluída en la modelización. Para los dos tipos de rechazo el comportamiento es el mismo (Figura 1a): en tamaños pequeños apenas divergen las probabilidades de éxito para los dos valores de τ^2, si bien a partir de $n_r = 20$ resulta ya apreciable el efecto de la restricción inicial más fuerte sobre la proximidad entre θ_o y θ_r ($\tau^2 = 1/5$), que da a la réplica mayor probabilidad de reproducir las conclusiones de ε_o.

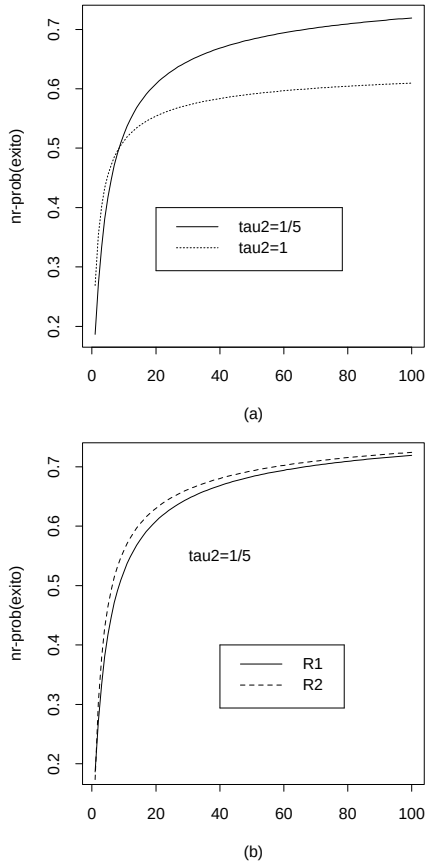

Figura 1. Reproducción de Conclusiones. Éxito $\equiv$ rechazo de $H_0^r : \theta_r \leq 0$. Probabilidad de éxito en función del tamaño de la réplica.

(1a) Rechazo con p-valor —R1—, para estudios próximos ($\tau^2 = 1/5$) y alejados ($\tau^2 = 1$) .

(1b) Estudios próximos. Rechazos R1 —réplica aislada— y R2 —combinando original y réplica—.

3.3. Comparación de poblaciones

Un investigador podría estar interesado en repetir una experiencia estadística previa sobre cierta población de interés, distinta en principio a aquella sobre la que se muestreó en el estudio original, con la pretensión de comparar ambas poblaciones respecto del efecto estudiado. Este sería el contexto de las investigaciones basadas en estudiar la evolución (en el tiempo) de una determinada población, o su sensi-

bilidad ante modificaciones en las condiciones experimentales. Una discrepancia a priori grande entre las poblaciones muestreadas en los dos estudios vendrá dada por un valor «grande» para τ^2 (que representaba la variabilidad inicial de θ_o y θ_r).

Consideraremos a partir de ahora que tanto en ε_o como en la réplica, las inferencias se llevan a cabo (por parte del investigador) desde una perspectiva totalmente bayesiana. Así, el resultado del primer estudio (x_o) condicionará siempre la inferencia en la réplica (toda función de x_r dependerá implícitamente de x_o: $\gamma(x_r) \equiv \gamma(x_r|x_o)$).

Un objetivo natural que se plantea el investigador cuando replica en esta situación consiste en inferir sobre la diferencia de medias $\delta = \theta_r - \theta_o$. Una vez observada la réplica, la inferencia sobre las medias θ_o y θ_r se lleva a cabo con la distribución final:

$$p(\theta_o, \theta_r|x_o, x_r) \propto f(x_o, x_r|\theta_o, \theta_r) \cdot p(\theta_o, \theta_r),$$

con

$$p(\theta_o, \theta_r) = \int p(\theta_o, \theta_r|\mu) \cdot p(\mu)\, d\mu \propto p(\theta_r|\theta_o),$$

de donde se obtiene la distribución a posteriori para la diferencia δ:

$$p(\delta|x_o, x_r) \quad = \quad N(\delta|dQR, \gamma_o^2 Q + \gamma_r^2 Q^2 R), \tag{8}$$

con $d = x_r - x_o$,

$$Q = \frac{2\tau^2}{\gamma_o^2 + 2\tau^2}, \quad \text{y} \quad R = \frac{\gamma_o^2 + 2\tau^2}{\gamma_o^2 + \gamma_r^2 + 2\tau^2}.$$

Los objetivos en la réplica respecto a la inferencia a realizar sobre δ podrían ser:

C1. conseguir una estimación de δ lo suficientemente precisa,
C2. resolver el contraste de igualdad de efectos ($\theta_o = \theta_r$).

3.3.1. Precisión

Una forma de cuantificar la precisión de las inferencias a posteriori sobre la diferencia de medias consiste en utilizar la longitud de una región creíble para δ, obtenida a partir de su distribución final. Una réplica en las condiciones expuestas en 3.3 se considerará un éxito si dicha longitud, $L_\delta(x_r, n_r)$, resultara suficientemente pequeña (inferior a cierto S fijado por el investigador):

$$L_\delta(X_r, n_r) \;\leq\; S. \tag{9}$$

El experimentador elegiría para replicar aquel tamaño n_r que le proporcionara suficientes garantías de éxito:

$$Prob(\text{ éxito }) = P^{X_r|x_o}[L_\delta(X_r, n_r) \leq S] \geq M.$$

En nuestro modelo, la longitud de la región creíble para δ, de contenido probabilístico $1-\alpha$, resulta independiente de X_r:

$$L_\delta(X_r, n_r) \equiv L_\delta(n_r) = 2 \cdot z_{1-\alpha/2} \cdot \sqrt{v_\delta^2(n_r)},$$

donde $z_{1-\alpha/2}$ es el cuantil de la normal estándar y v_δ^2 la varianza de la distribución a posteriori de δ, que sólo depende del tamaño n_r de la réplica. Así, con probabilidad 1 el investigador podrá inferir sobre δ con una precisión dada, $1/v_\delta^2(n_r)$, siempre que replique con un tamaño n_r tal que:

$$v_\delta^2(n_r) \leq S^*,$$

para un S^* convenientemente pequeño.

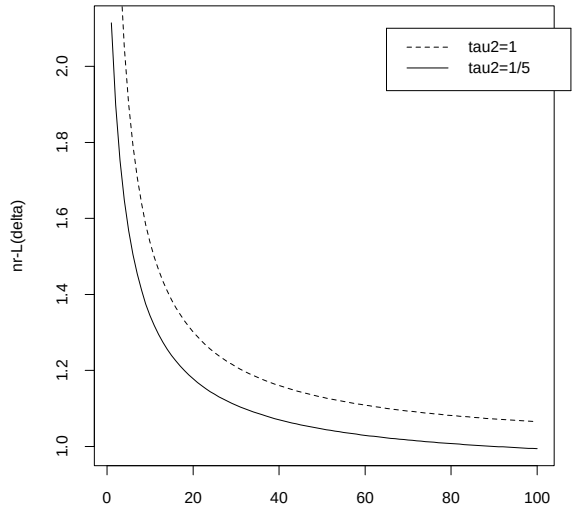

Figura 2. Comparación de Poblaciones. Éxito $\equiv$ Precisión al inferir sobre $\delta = \theta_r - \theta_o$. Longitud del intervalo creíble para δ (de probabilidad 0.95), L_δ, en función de n_r (estudios próximos y alejados).

En la Figura 2 hemos representado la longitud de la región creíble al 95% para δ, $L_\delta(n_r)$, en los casos $\tau^2 = 1$ y $\tau^2 = 1/5$. Una mayor proximidad inicial entre los estudios original y réplica repercute (como era de suponer) en una mayor precisión en la inferencia final sobre δ. En ambos supuestos iniciales el comportamiento es similar: un decrecimiento rápido de la longitud de la región creíble, seguido de una estabilización de su valor (a partir de $n_r = 50$) por debajo de 1.17. Ni una mayor

cercanía a priori entre las poblaciones, ni la incorporación de más observaciones cuando partamos de poblaciones supuestamente distintas, proporcionarán ya una ganancia apreciable de precisión en la inferencia sobre δ.

3.3.2. Contraste de igualdad

Una forma de inferir acerca de la discrepancia entre ambas poblaciones (original y réplica) es plantear el contraste sobre la igualdad de sus medias:

$$\begin{cases} H_0^\delta & : \quad \delta = 0 \\ H_1^\delta & : \quad \delta \neq 0, \end{cases}$$

donde $\delta = \theta_r - \theta_o$.

Desde una perspectiva clásica, suele considerarse «éxito» en un contraste, el rechazo de la hipótesis nula. Aquí llevaremos a cabo el análisis desde una perspectiva bayesiana y nuestro único objetivo será el de resolver el contraste, ya sea a favor de una o de otra hipótesis, con «suficiente» grado de confianza. El éxito en la réplica consistirá pues, en que con su resultado podamos discriminar fácilmente entre ambas hipótesis en base a sus probabilidades finales, o lo que es igual, en obtener un factor Bayes a favor de una de dichas hipótesis lo bastante grande.

El factor Bayes (el cociente entre los «odds» a posteriori y los «odds» a priori) es una herramienta bayesiana que cuantifica la evidencia a favor o en contra de las hipótesis del contraste. Puesto que no involucra explícitamente las probabilidades iniciales para cada hipótesis,

$$\frac{P(H_0)}{P(H_1)} \cdot B_{01} = \frac{P(H_0|datos)}{P(H_1|datos)},$$

puede interpretarse como un cociente de las verosimilitudes ponderadas:

$$\begin{aligned} B_{01}^\delta &= \frac{f(d|H_0^\delta)}{f(d|H_1^\delta)} \\ &= \frac{N(d|0,\gamma_o^2+\gamma_r^2)}{N(d|0,\gamma_o^2+\gamma_r^2+2\tau^2)}, \end{aligned} \tag{10}$$

donde:

$$d = x_r - x_o, \text{ y}$$
$$f(d|H_1^\delta) = \int_{\delta\neq 0} f(d|\delta)\cdot p(\delta)\ d\delta.$$

Un factor Bayes B_{01} grande indica evidencia a favor de H_0 (y pequeño, a favor de la alternativa). Fijado el nivel de confianza ($B > 1$) que exigimos para concluir sobre el contraste planteado, éste se resolverá a dicho nivel cuando:

$$\begin{array}{lcl} \text{se acepte } H_0 & \Leftrightarrow & B_{01} > B \\ & \text{ó} & \\ \text{se acepte } H_1 & \Leftrightarrow & B_{10} > B \Leftrightarrow B_{01} \leq 1/B, \end{array}$$

siendo B_{10} el índice de evidencia a favor de la hipótesis alternativa.

Es decir, el ***éxito*** vendrá definido por obtener con la replicación:

$$B_{01}^{\delta} > B \quad \text{ó} \quad B_{01}^{\delta} \leq 1/B. \tag{11}$$

Al depender B_{01}^{δ} de X_r, para un n_r fijo la probabilidad de éxito se obtendrá integrando la distribución predictiva del resultado de la réplica sobre el rango de valores de X_r que satisfacen la condición [11], esto es, sobre la región:

$$R_B^{\delta}(X_r, n_r) = \{x_r : B_{01}^{\delta}(x_r, n_r) > B \text{ ó } B_{01}^{\delta}(x_r, n_r) \leq 1/B\}.$$

Obviamente, cuanto mayor sea el nivel de evidencia que exijamos para concluir a favor de una de las hipótesis, más reducida será la región de éxito y por tanto menor la probabilidad de resolver el contraste.

La planificación «óptima» de la réplica resultará del mínimo n_r para el que se obtengan las garantías de éxito (M) exigidas por el experimentador:

$$Prob(\text{ éxito }) = \int_{R_B^{\delta}(x_r, n_r)} m(x_r|x_o)\, dx_r \geq M.$$

Al pretender concluir sobre un contraste con hipótesis nula ***puntual***, es de esperar que en su resolución se aprecie una sensibilidad importante respecto de la información a priori utilizada. En efecto, si asumimos a priori una discrepancia «leve» entre las poblaciones original y réplica ($\tau^2 = 1/5$), serán precisos muchos datos (o muy «raros») para conseguir evidencia suficiente en contra de H_0. La probabilidad de éxito en este caso se espera pequeña, como de hecho resulta (inferior a 0.7, aun con 300 datos). Con $\tau^2 = 1$, caso en el que ya a priori asumimos cierta discrepancia entre la población inicial y la replicada, en principio se espera poder discriminar mejor ambas poblaciones, al tener de partida más peso para la hipótesis «más grande» (la alternativa a la puntual). No son necesarios entonces muchos datos para lograr concluir sobre el contraste, aun en el caso más restrictivo con B=4 (ver Tabla 1). Cuando sólo se exige (para concluir sobre el contraste) el doble de evidencia a favor de alguna de las hipótesis (B=2), la probabilidad de éxito resulta mayor que 0.85 tan sólo duplicando el tamaño original en la réplica ($n_r = 30$).

Tabla 1. *Comparación de Poblaciones. Prob($B_{01}^{\delta} < 1/B$ ó $B_{01}^{\delta} > B$)*

ÉXITO: Resolver $\theta_r = \theta_o$ versus $\theta_r \neq \theta_o$

	n_r	15	30	100	300
	$\tau^2 = 1/5$	0.3364	0.5735	0.6613	0.6846
B=2					
	$\tau^2 = 1$	0.8374	0.8645	0.8851	0.8914
B=4	$\tau^2 = 1$	0.6366	0.7633	0.8129	0.8262

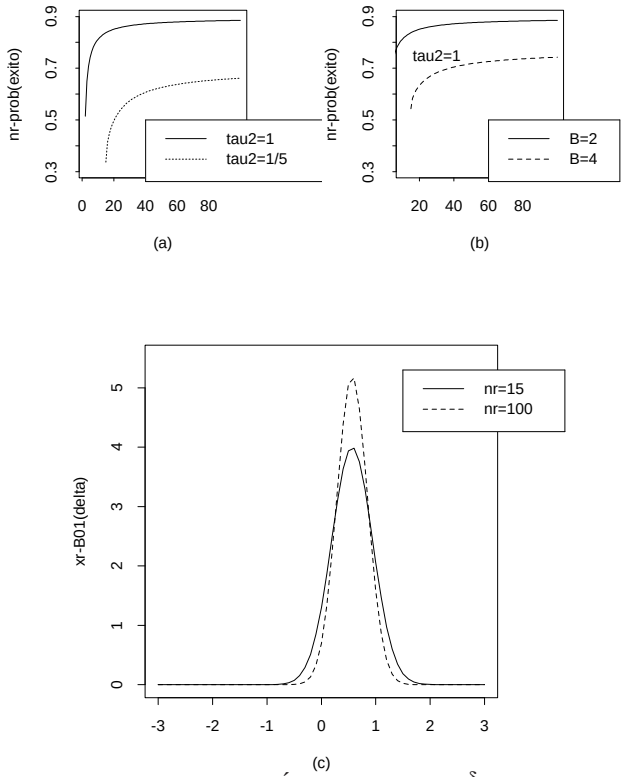

Figura 3. Comparación de Poblaciones. Éxito $\equiv$ resolver $H_0^{\delta} : \theta_r = \theta_o$ con el factor Bayes B_{01}^{δ}.

(3a) Éxito $\equiv \boldsymbol{B}_{01}^{\delta} < 1/2$ ó $B_{01}^{\delta} > 2$. Probabilidad de éxito versus n_r para estudios próximos y alejados.

(3b) Éxito $\equiv \boldsymbol{B}_{01}^{\delta} < 1/B$ ó $B_{01}^{\delta} > B$. Estudios alejados. Probabilidad de éxito versus n_r para B=2 y B=4.

(3c) Estudios alejados. $B_{01}^{\delta}(x_r)$ versus x_r para $n_r = 15$ y $n_r = 100$.

Dadas las restricciones de nuestro modelo y los valores asumidos a priori para los parámetros, la región de éxito, dependiente de n_r y B, resulta en ocasiones «imposible». Cuando τ^2 es pequeño, valores altos de B y bajos de n_r dan lugar a regiones imposibles (dicha condición de «posibilidad» está detallada en el apéndice 1). Es el caso de $\tau^2 = 1/5$ y $B = 4$.

En la Figura 3a está representada la probabilidad de éxito en función de n_r para ambas situaciones (considerablemente menor en $\tau^2 = 1/5$ que en $\tau^2 = 1$), definiendo la región de éxito con $B = 2$. En la Figura 3b apreciamos las diferencias en la amplitud de dicha región cuando tomamos $B = 2$ y $B = 4$ para resolver el contraste.

En la Figura 3c hemos representado la magnitud del factor Bayes en función del resultado a observar en la réplica, para dos tamaños distintos ($n_r = 15$ y $n_r = 100$). El rango de valores de X_r para los que se consigue evidencia a favor de H_0^r no tiene amplitud mayor a 2 ya en el caso más desfavorable (en el que la réplica cuenta con menos observaciones —$n_r = 15$—), disminuyendo a medida que aumenta n_r. Más datos proporcionarán inferencias más precisas y permitirán discriminar mejor entre las hipótesis planteadas (un x_r próximo a x_o otorgará más evidencia a H_0^δ cuando $n_r = 100$ que cuando $n_r = 15$).

4. DETECCIÓN DE SESGO CON LA RÉPLICA

Es frecuente encontrar primeros análisis llevados a cabo con un diseño deficiente, a veces incluso inexistente, que puede haber dado lugar a la introducción de sesgo en el efecto percibido en la muestra. La replicación permite repetir la experiencia controlando factores y/o variables que podrían haber quedado «sueltos» en dicho primer estudio, reduciendo o eliminando así el sesgo. El éxito en este tipo de réplicas estará basado en la inferencia sobre el sesgo.

Una forma posible de modelizar el sesgo β en el experimento original (y su ausencia en ε_r) consiste en suponer para las medias poblacionales la relación:

$$\begin{aligned} \theta_o &= \mu + \beta \\ \theta_r &= \mu. \end{aligned}$$

El modelo jerárquico M3 incorpora la información a priori sobre el sesgo β a la modelización de las restantes variables.

4.1. Modelo M3

I. DATOS

Con muestras condicionalmente independientes, suponemos que la distribución de los estadísticos suficientes es:

$$f(x_o|\mu,\beta) = N(x_o|\mu+\beta,\gamma_o^2), \tag{12}$$

$$f(x_r|\mu) = N(x_r|\mu,\gamma_r^2), \tag{13}$$

con $\gamma_i^2 = \sigma^2/n_i$ $(i = o,\ r)$, y σ^2 conocida (=1).

II. PARÁMETROS

Modelizamos con una distribución mínimo-informativa la información a priori sobre la media μ. Para el sesgo de ε_o asumimos una distribución inicial normal centrada en cero y con varianza conocida Δ^2, mayor o menor en función de las sospechas iniciales sobre su existencia:

$$p(\mu) \propto \text{cte}\ , \tag{14}$$

$$p(\beta) = N(\beta|0,\Delta^2). \tag{15}$$

En nuestra inferencia hemos considerado dos valores para Δ^2 relativamente distintos:

$\Delta^2 = 1$: Suponemos que el sesgo, en caso de existir, no puede ser demasiado grande.

$\Delta^2 = 5$: La información a priori sobre la existencia de sesgo en ε_o es débil; estaríamos dando probabilidad apreciable a magnitudes grandes para el sesgo.

El análisis de sensibilidad se obtendrá, en esta primera aproximación, de la comparación entre los resultados para estos dos valores. Con posterioridad (modelo M3A) se supondrá Δ^2 aleatorio.

Para calcular la probabilidad de éxito en la réplica utilizaremos, como ya hicimos en los otros tipos de réplicas, la distribución predictiva para su resultado, obtenida de la combinación de información:

$$\begin{aligned} m(x_r|x_o) &= \int f(x_r|\mu)\cdot p(\mu|x_o)\, d\mu \\ &= N(x_r|x_o,\gamma_o^2+\gamma_r^2+\Delta^2), \end{aligned} \tag{16}$$

donde

$$p(\mu|x_o) = \int p(\mu|\beta, x_o) \cdot p(\beta)\, d\beta,$$

dado que, en nuestro caso,

$$p(\beta|x_o) \propto f(x_o|\beta) \cdot p(\beta) \propto p(\beta).$$

4.2. Inferir sobre el sesgo

Proponemos la inferencia sobre el contraste de existencia de sesgo en el primer estudio:

$$\begin{cases} H_0^\beta & : \quad \beta = 0 \\ H_1^\beta & : \quad \beta \neq 0, \end{cases}$$

utilizando de nuevo el factor Bayes. Concluiremos sobre dicho contraste cuando obtengamos un valor para él lo suficientemente grande (a favor de H_0^β), o lo suficientemente pequeño (a favor de H_1^β).

Siendo:

$$B_{01}^\beta = \frac{N(x_r|x_o, \gamma_o^2 + \gamma_r^2)}{N(x_r|x_o, \gamma_o^2 + \gamma_r^2 + \Delta^2)}, \tag{17}$$

la región de éxito vendrá dada (para un n_r fijo) por:

$$R_B^\beta(X_r, n_r) = \{x_r \,:\, B_{01}^\beta(x_r, n_r) > B \;\text{ ó }\; B_{01}^\beta(x_r, n_r) \leq 1/B\},$$

donde B representa el grado de confianza exigido para resolver el contraste.

Tabla 2. *Inferencia sobre el Sesgo. Prob($B_{01}^\beta \leq 1/B$ ó $B_{01}^\beta \geq B$)*

ÉXITO: Resolver $\beta = 0$ versus $\beta \neq 0$.

	n_r	15	30	100	300
B=2	$\Delta^2 = 1$	0.7417	0.7880	0.8220	0.8322
	$\Delta^2 = 5$	0.9081	0.9228	0.9342	0.9377
B=3	$\Delta^2 = 1$	#	0.6105	0.6890	0.7098
	$\Delta^2 = 5$	0.8490	0.8740	0.8931	0.8989
B=4	$\Delta^2 = 1$	#	#	#	#
	$\Delta^2 = 5$	0.8006	0.8352	0.8611	0.8689

La probabilidad de éxito para cada tamaño n_r se calcula integrando la predictiva $m(x_r|x_o)$ sobre dicha región, y de la condición [1] se obtendría el *mejor* tamaño con que replicar.

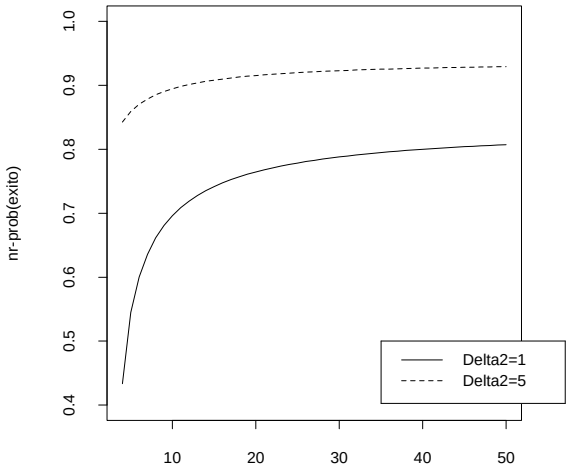

Figura 4. Detección de Sesgo. Éxito $\equiv$ resolver $H_0^{\beta} : \beta = 0$ con $B_{01}^{\beta} < 1/2$ ó $B_{01}^{\beta} > 2$.: Probabilidad de éxito versus n_r, asumiendo distinta información inicial sobre el sesgo de ε_o

En la Figura 4 está representada la probabilidad de éxito en función de n_r cuando el investigador está dispuesto a aceptar aquella hipótesis para la que obtenga el doble de evidencia (B=2). La diferencia entre las probabilidades de éxito con $\Delta^2 = 1$ y $\Delta^2 = 5$ es importante, si bien resultaba esperable (puesto que con $\Delta^2 = 5$ estamos contemplando la posibilidad de mayores sesgos, es razonable que esperemos poder detectarlo más fácilmente). Este efecto queda reflejado en la Tabla 2. Con $n_r = 30$ y $\Delta^2 = 1$, si bien para $B = 2$ la probabilidad de éxito está próxima a 0.8, para B=3 se queda sólo en 0.6. Para $\Delta^2 = 5$ la probabilidad de resolver el contraste con cualquiera de los tres tamaños es muy alta (con B=4, superior a 0.8). Valores pequeños de Δ^2, niveles de confianza (B) altos y tamaños de la réplica (n_r) bajos dan lugar a regiones «imposibles» (ver apéndice 1). En la Tabla 2 el símbolo # designa incompatibilidades de ese tipo.

5. REPRODUCCIÓN Y COMPARACIÓN: τ^2 DESCONOCIDA

Como una extensión del modelo M12 que proponíamos para analizar la réplica cuando se pretendía reproducir las conclusiones de un estudio previo (sec. 3.2) o comparar la población muestreada en la réplica con la de cierto experimento original (sec. 3.3), modelizamos las relaciones entre los parámetros asumiendo un mayor desconocimiento inicial.

Conservando la estructura jerárquica de M12, las medias poblacionales θ_o y θ_r se asumen intercambiables, normales, con media μ y varianza τ^2, ambas desconocidas. En el tercer nivel de dicho modelo será entonces necesario precisar la información inicial sobre el nuevo parámetro introducido (τ^2). Al tratarse de una medida de la distancia entre las poblaciones de ε_o y ε_r, se asume (es lo usual) una distribución a priori gamma invertida:

$$p(\tau^2) = GaI(\tau^2|a+1, ak). \tag{18}$$

Esta modelización es equivalente a asumir que, a priori y condicionando a μ, las medias poblacionales θ_o y θ_r son intercambiables (en lugar de independientes como en M12), con distribuciones marginales que son *t-Student* (las colas son algo mayores que las del modelo normal de M12):

$$p(\theta_i|\mu) = St\left(\theta_i|\mu, \frac{ak}{a-1}, 2(a+1)\right), \quad i = o, r.$$

Así k, el valor esperado para la variabilidad a priori de las medias, será un indicativo de la discrepancia asumida inicialmente entre la población de ε_o y la de la réplica. Al igual que en M12, consideraremos en nuestro ejemplo los valores $k = 1/5$ para una discrepancia a priori pequeña entre ambas poblaciones (S1), y $k = 1$ para cuando asumamos cierto «alejamiento» (S2).

El parámetro «a» da idea sobre la certidumbre de nuestra información inicial acerca de la distribución de las medias θ_o y θ_r. Un valor de $a = 4$ proporciona unas colas para la distribución Student a priori no demasiado altas (10 grados de libertad), de forma que el alejamiento de la modelización normal de M12 no resulte excesivo.

La modificación introducida en el modelo da lugar a una complicación considerable en el cálculo de la distribución predictiva, que se obtiene de una integral no analítica:

$$m(x_r|x_o) = \int m(x_r|x_o, \tau^2) \cdot p(\tau^2)\, d\tau^2, \tag{19}$$

donde $m(x_r|x_o,\tau^2)$ es la predictiva [5] resultante en M12:

$$m(x_r|x_o,\tau^2) = N(x_r|x_o,\gamma_o^2+\gamma_r^2+2\tau^2).$$

(La expresión [19] de la predictiva refleja el hecho de que $p(\tau^2|x_o) = p(\tau^2)$).

Problemas de cálculo surgen también para las restantes densidades a posteriori y factores Bayes con los que se lleva a cabo el análisis. Utilizamos la simulación, la integración numérica y la aproximación Monte-Carlo para obtener estimaciones con qué evaluar la probabilidad de éxito al replicar. Las densidades de interés son aproximadas por estimadores *kernel* (ver Silverman, 1986) que suavizan los histogramas confeccionados con las simulaciones obtenidas.

Dada la complicación del cálculo, el análisis de la replicación lo reducimos considerando sólo tres valores de n_r: 15 (igual al tamaño de ε_o), 30 (doble) y 100 (considerablemente mayor).

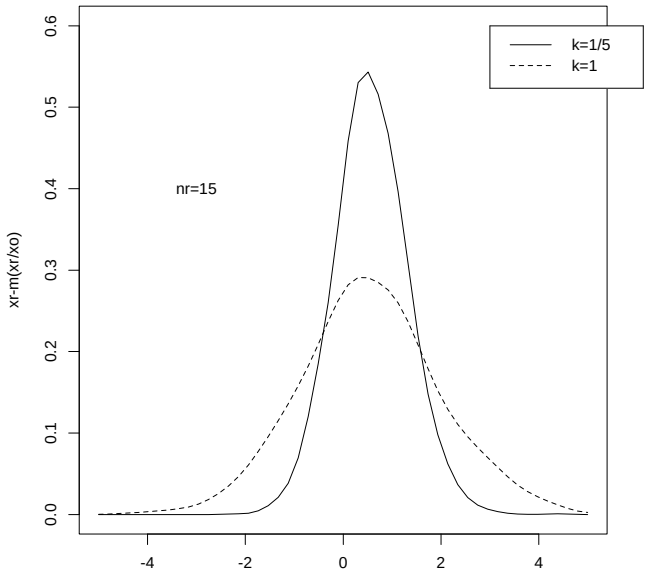

Figura 5. Predicción del resultado de la réplica con τ^2 desconocida. Estudios próximos: $k = 1/5$. Estudios alejados: $k = 1$.

En la Figura 5 hemos representado la aproximación que obtenemos de la densidad predictiva de X_r para cada uno de los tres tamaños muestrales de interés y las dos

posibilidades de k (1/5 y 1). Cuando la divergencia a priori entre las poblaciones es menor (en valor esperado, esto es, $k = 1/5$), la predictiva para X_r resulta más precisa (con menor varianza) y apuntada que en el caso en que inicialmente asumimos mayor discrepancia ($k = 1$). Dicha distribución aparece centrada alrededor de la observación del primer estudio (≈ 0.5). El apuntamiento y la precisión crecen con el tamaño de la réplica (n_r), como era de esperar.

Vayamos a cada uno de los supuestos que ya consideramos con M12.

5.1. Reproducción de conclusiones

El problema es idéntico al planteado en la sección 3.2, con sus dos tipos de éxito:

R1: $p_r \leq 0.05$,
R2: $P(\theta_r \leq 0 | x_o, x_r) \leq 0.05$.

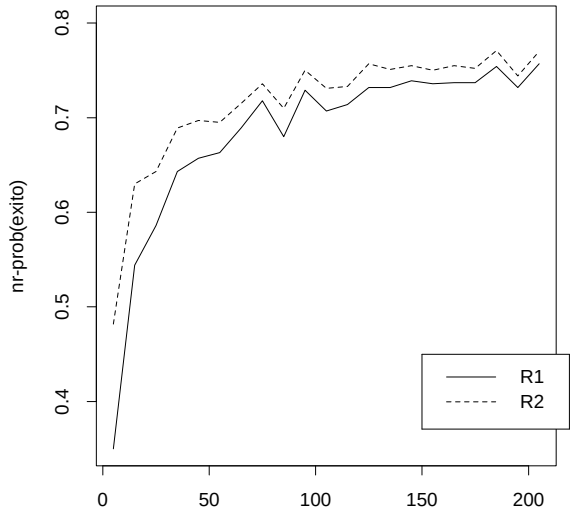

Figura 6. Reproducción de Conclusiones con τ^2 desconocida. Éxito $\equiv$ rechazo de H_0^r : $\theta_r \leq 0$. Estudios próximos. Probabilidad de éxito versus n_r. Rechazos R1 (con p-valor) y R2 (con probabilidad a posteriori para H_0^r).

Las dos formas de rechazo nos conducían a sendas regiones de éxito cuya probabilidad evaluábamos integrando respecto de la predictiva $m(x_r|x_o)$. Ahora, al no

disponer de su expresión analítica, estimamos las probabilidades de éxito mediante ***conteos*** a partir de simulaciones de la distribución predictiva. (Al tomar 2000 simulaciones, el error estándar del estimador resulta inferior a 0.01).

El comportamiento es similar al observado en el análisis de M12 (ver Figura 6). La probabilidad de rechazar combinando información (R2) resulta en todo momento, mayor que la de rechazar sólo con la réplica (R1), y la estabilización (por debajo de 0.8) es apreciable a partir de $n_r = 100$.

5.2. Comparación de poblaciones

Como en la sección 3.3, se pretende inferir sobre la diferencia de medias $\delta = \theta_r - \theta_o$, para lo que se proponen dos tipos de éxito:

C1. conseguir una estimación lo suficientemente precisa;
C2. resolver el contraste $H_0^{\delta} : \delta = 0$ con suficiente grado de confianza.

Asumimos para nuestro ejemplo $k = 1$ (poblaciones distantes «a priori»).

5.2.1. Precisión

Al no disponer de una expresión analítica para la distribución a posteriori de la diferencia de medias δ:

$$p(\delta|x_o, x_r) = \int p(\delta|x_o, x_r, \tau^2) \cdot p(\tau^2|x_o, x_r)\, d\tau^2, \tag{20}$$

donde

$$p(\tau^2|x_o, x_r) \propto m(x_r|x_o, \tau^2) \cdot p(\tau^2),$$

no tenemos una expresión «cerrada» para la varianza a posteriori de δ, $v_{\delta}^2(X_r, n_r)$, en función de la cual podíamos concluir sobre la precisión que proporcionaba la réplica para inferir sobre la diferencia de medias. Sin embargo, dado que dicha varianza es aleatoria (depende de la variable no observada X_r), fijando n_r obtenemos una muestra «aproximada» de su distribución predictiva (por estimación Monte-Carlo —ver apéndice 2—). Con la densidad predictiva de la varianza a posteriori se estima la probabilidad de éxito, definido este por:

$$v_{\delta}^2(X_r, n_r) \leq S^*,$$

evaluando el área encerrada a la izquierda de S^*, o simplemente como ya hacíamos con R1 y R2, utilizando conteos de las estimaciones de la muestra por debajo de S^*. La «mejor» planificación de la réplica se obtendrá de aquel tamaño n_r que proporcione suficientes garantías de éxito.

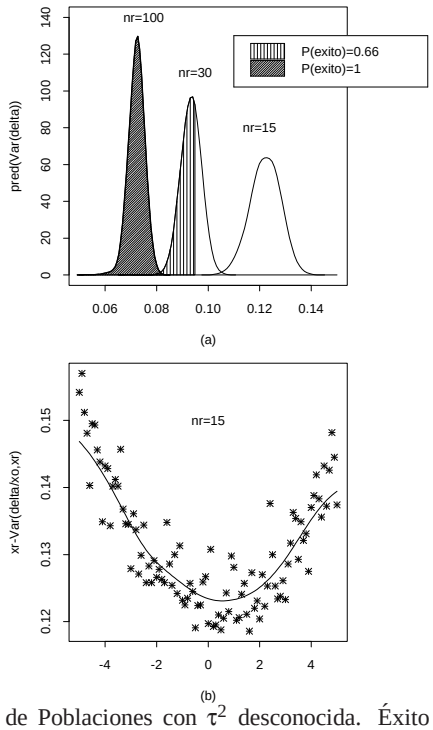

Figura 7. Comparación de Poblaciones con τ^2 desconocida. Éxito $\equiv$ Precisión al inferir sobre $\delta = \theta_r - \theta_o$. Estudios alejados.

(7a) Éxito: $Var(\delta|x_o, X_r) \leq 0.095$. Predicción de $Var(\delta|x_o, X_r)$ y probabilidad de éxito para $n_r = 15, 30$ y 100.

(7b) Aproximación MC de $Var(\delta|x_o, x_r)$ versus x_r.

En la Figura 7a representamos conjuntamente las densidades predictivas de la varianza a posteriori de δ (aproximación kernel) para los tres tamaños de interés. La influencia del tamaño de la réplica es importante: las tres distribuciones representadas ($n_r = 15$, $n_r = 30$ y $n_r = 100$) no se solapan apenas. Cuanto mayor es n_r, mayor es el apuntamiento de esta distribución y más próxima a cero queda su media. Tomando $S^* = 0.095$, la probabilidad de éxito (área sombreada) planificando con $n_r = 15$ es nula; con $n_r = 30$ resulta ya de 0.66, y con $n_r = 100$ es máxima (todos los valores «probables» de la varianza a posteriori quedan por debajo de 0.095).

Comparamos en la Figura 7b la magnitud de la varianza $v^2_\delta(x_r, 15)$ (su aproximación) con el resultado de la réplica (x_r), ajustando una curva a la representación de los puntos. Esta varianza resulta mínima cuando el valor x_r de la réplica está próximo al observado en el primer estudio, creciendo a medida que éste es más «raro» respecto

de x_o. De lo observado con otros valores de n_r, cuanto mayor es el tamaño de la réplica mayor resulta la precisión con que se infiere sobre δ (menor es la magnitud de su varianza a posteriori), si bien el comportamiento respecto de x_r es similar.

5.2.2. Contraste de igualdad

Al utilizar el factor Bayes de la diferencia de medias B_{01}^{δ} para concluir sobre el contraste de igualdad de poblaciones, de nuevo nos encontramos con dificultades analíticas:

$$B_{01}^{\delta}(x_r, n_r) = \frac{N(x_r|x_o, \gamma_o^2 + \gamma_r^2)}{\int m(x_r|x_o, \tau^2) \cdot p(\tau^2) d\tau^2}.$$

Fijando el tamaño muestral en la réplica (n_r) obtenemos una muestra de la distribución predictiva para B_{01}^{δ} utilizando integración numérica. Puesto que el éxito lo definíamos [11] por resolver el contraste a favor de una u otra hipótesis a un nivel de confianza B, la probabilidad de éxito es estimada mediante conteos a partir de la muestra obte-
nida.

Representamos en la Figura 8a la distribución predictiva para B_{01}^{δ} considerando $n_r = 15$ y 100. El comportamiento es similar para ambos tamaños. La densidad predictiva del factor Bayes resulta bimodal, con la moda máxima alrededor de cero (evidencias en contra de H_0^{δ}) y la menor (evidencias a favor de la hipótesis puntual) desplazándose hacia la derecha a medida que crece el tamaño de la réplica. Con n_r aumenta la probabilidad de concluir ***con mayor confianza*** sobre el contraste, esto es, de discriminar mejor entre las hipótesis. En la gráfica sombreamos la región complementaria a la región de éxito, definido éste por conseguir para alguna de las hipótesis un grado de evidencia cuatro veces superior al conseguido para su alternativa. Con B=4, una réplica de 15 observaciones da una probabilidad muy baja a lograr concluir sobre el contraste; ya con 100 observaciones es posible discriminar entre las dos hipótesis del contraste. El efecto del tamaño muestral sobre la probabilidad de éxito es más apreciable cuando el grado de exigencia para resolver el contraste (B) es mayor (ver Tabla 3).

Tabla 3. *Comparación de Poblaciones. Prob($B_{01}^{\delta} \geq B$ ó $B_{01}^{\delta} \leq 1/B$)*

ÉXITO: Resolver $\theta_r = \theta_o$ versus $\theta_r \neq \theta_o$			
n_r	15	30	100
B=2	0.797	0.8525	0.879
B=4	0.596	0.712	0.7935

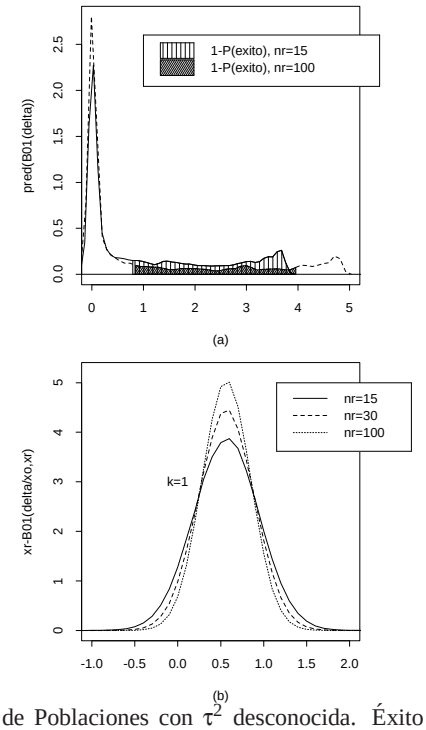

Figura 8. Comparación de Poblaciones con τ^2 desconocida. Éxito $\equiv$ resolver $H_0^\delta : \delta = 0$ con B_{01}^δ. Estudios alejados.

(8a) Éxito $\equiv$ $B_{01}^\delta < 1/4$ ó $B_{01}^\delta > 4$. Predicción de B_{01}^δ y probabilidad de éxito para $n_r = 15$ y 100.

(8b) $B_{01}^\delta(x_r)$ versus x_r para $n_r = 15, 30$ y 100.

Hemos representado también la magnitud del factor Bayes para la diferencia de medias en función del valor observado para X_r (Figura 8b). Para valores de X_r próximos a x_o (evidencias a favor de H_0^δ), el factor Bayes B_{01}^δ crece con n_r. Valores de X_r alejados de x_o inducen mayores evidencias a favor de la alternativa y en contra de la igualdad de poblaciones (B_{01}^δ crece) cuando el tamaño de la réplica es más grande.

6. DETECCIÓN DE SESGO: δ^2 DESCONOCIDA

Buscando también robustecer nuestros resultados en el análisis de la réplica cuando el objetivo de ésta consiste en detectar sesgo en un estudio previo (o conseguir evidencia en contra de su existencia), añadimos a la modelización jerárquica M3, un nivel más.

6.1. Modelo M3A

En dicha modelización asumíamos para el sesgo β una distribución a priori normal, centrada en cero y con varianza Δ^2 conocida. Ahora consideramos esta varianza desconocida, por lo que surge un tercer nivel en la jerarquía destinado a la distribución a priori para Δ^2 (una gamma invertida):

$$p(\Delta^2) = GaI(\Delta^2|b+1, bk). \tag{21}$$

Esto es equivalente a asumir como distribución inicial para el sesgo β, una distribución *t-Student*:

$$p(\beta) = St\left(\beta|0, \frac{bk}{b+1}, 2(b+1)\right). \tag{22}$$

Siendo b el parámetro que da la amplitud de las colas en la distribución [22], tomamos $b = 1$ (que le da 4 g.l.$\equiv$ colas altas) asumiendo así muy poca certidumbre a priori sobre la existencia de sesgo en ε_o. Como valores esperados para la varianza inicial del sesgo consideramos, siguiendo M3, $k = 1$ y $k = 5$, que informan sobre su posible magnitud.

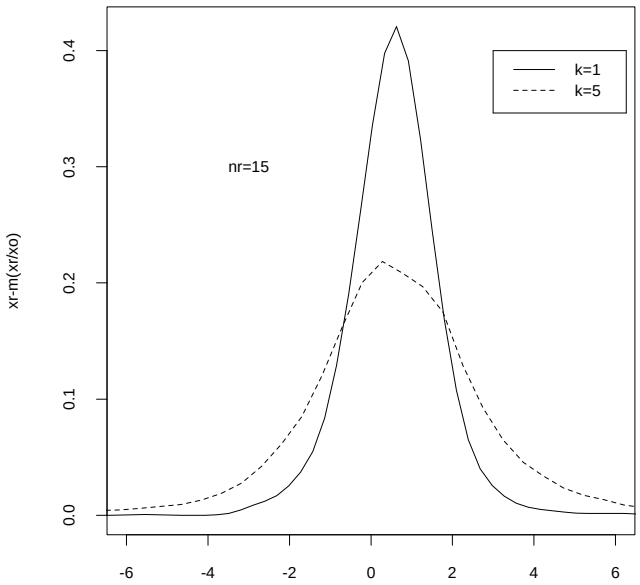

Figura 9. Predicción del resultado de la réplica asumiendo sesgo en ε_o y Δ^2 desconocida. Información precisa sobre el sesgo: $k = 1$. Información dispersa sobre β: $k = 5$.

La distribución predictiva que obtenemos ahora para llevar a cabo la planificación de la réplica resulta:

$$m(x_r|x_o) = \int m(x_r|x_o,\Delta^2)\cdot p(\Delta^2)\ d\Delta^2, \tag{23}$$

donde $m(x_r|x_o,\Delta^2)$ es la predictiva [16] obtenida en M3.

Al igual que en el modelo M12A, utilizamos la integración numérica para resolver las integrales no analíticas (caso del factor Bayes, que involucra integraciones unidimensionales) y la simulación y aproximación kernel para obtener las distribuciones de interés.

En la Figura 9 aproximamos la densidad predictiva de X_r con $n_r = 15$ para los dos valores de k, que afectan (como cabía esperar) al apuntamiento de la distribución: la variabilidad de la predicción de X_r resulta mayor cuanto menos sabemos a priori sobre el sesgo en ε_o (mayor k).

6.2. Inferir sobre el sesgo

Análogamente a como hacíamos en la sección 4.2, pretendemos resolver el contraste sobre la existencia de sesgo en el primer estudio con el factor Bayes a favor de la hipótesis nula de sesgo cero:

$$B_{01}^{\beta}(x_r,n_r) = \frac{N(x_r|x_o,\gamma_o^2+\gamma_r^2)}{m(x_r|x_o)}.$$

Fijado el índice de evidencia exigido para concluir, B, conseguiremos éxito cuando concluyamos a favor de alguna de las hipótesis:

$$\text{ÉXITO} \Leftrightarrow B_{01}^{\beta}(x_r,n_r) \geq B \text{ ó } B_{01}^{\beta}(x_r,n_r) \leq 1/B.$$

En la Figura 10 representamos las aproximaciones conseguidas para la densidad predictiva de $B_{01}^{\beta}(X_r)$ con $n_r = 15,\ 30$ y 100. El comportamiento es siempre el mismo: la distribución de $B_{01}^{\beta}(X_r)$ es bimodal, más variable cuanto mayor es el desconocimiento inicial sobre el sesgo (k grande). A mayor k más se apunta la densidad hacia la izquierda (reconocimiento de sesgo) y se achata y desplaza hacia la derecha (negación de sesgo). Cuanto menos sepamos inicialmente sobre el sesgo y mayor hayamos de asumir por tanto su variabilidad inicial, al dar cabida (con probabilidad apreciable) a valores grandes para el sesgo, más probabilidad tendremos de resolver el contraste a un nivel de confianza (B) alto, ya sea a favor de una u otra alternativa. Si la información inicial es más precisa ($k = 1$) sobre un sesgo muy pequeño (en torno a 0), el contraste será más difícil de resolver para un buen nivel de confianza, como se observa en la gráfica.

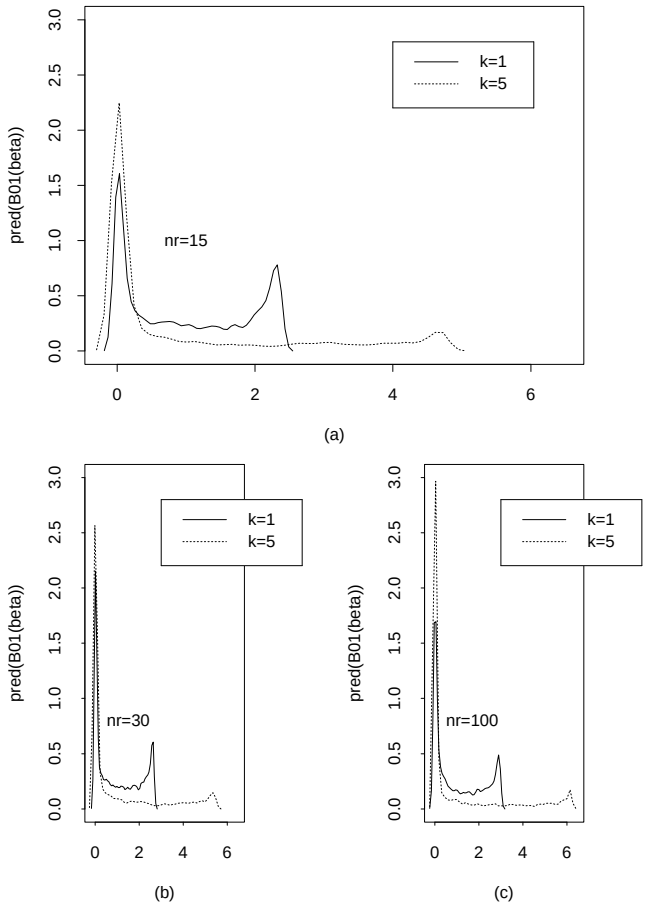

Figura 10. Detección de sesgo con Δ^2 desconocida. Predicción de B_{01}^{β} con sesgo disperso (k=5) y preciso (k=1) a priori.

(10a) $n_r = 15$.

(10b) $n_r = 30$.

(10c) $n_r = 100$.

Según se aprecia en la Figura 11, para k fijo el efecto del tamaño n_r es siempre el mismo: con n_r aumenta la variabilidad del factor Bayes y la magnitud de su moda más oriental, es decir, aumenta la probabilidad de conseguir evidencias altas a favor de la hipótesis nula de inexistencia de sesgo.

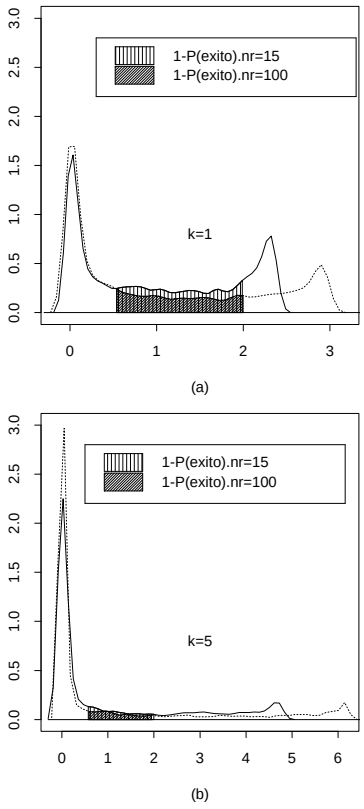

Figura 11. Detección de sesgo con Δ^2 desconocida. Éxito $\equiv$ resolver $H_0^\beta : \beta = 0$ con $B_{01}^\beta < 1/2$ ó $B_{01}^\beta > 2$. Predicción de B_{01}^β y probabilidad de éxito para $n_r = 15$ y $n_r = 100$.

(11a) Información a priori precisa sobre β.

(11b) Información a priori dispersa sobre β.

Tabla 4. *Inferencia sobre Sesgo en* ε_o. *Prob(*$B_{01}^\beta \leq 1/B$ *ó* $B_{01}^\beta \geq B$*)*

ÉXITO: Resolver $\beta = 0$ versus $\beta \neq 0$.

15	30	100		n_r		15	30	100
0.6493	0.6953	0.7559		B=2		0.8804	0.8764	0.9119
0.3692	0.4072	0.4462	k=1	B=3	k=5	0.7984	0.8074	0.8599
0.3417	0.3812	0.4202		B=4		0.7133	0.7454	0.8094

Fijado el grado de confianza B con el que vaya a resolverse el contraste, el tamaño de la réplica afecta más a la probabilidad de éxito cuanto mayor es la precisión inicial (k menor), como puede verse en la Tabla 4.

7. CONCLUSIONES Y EXTENSIONES

En este artículo hemos tratado de mostrar que es posible sistematizar, estudiar y planificar la replicación de una experiencia estadística previa antes de llevarla a cabo, en base a los distintos objetivos que pueden justificarla. Como objetivos sólo hemos considerado tres (los que suponíamos más comunes), y algunos métodos inferenciales, pero confiamos en haber transmitido el mensaje de que el análisis de la réplica es viable en cualquier situación (un estudio mucho más exhaustivo puede encontrarse en Mayoral, 1995).

Los modelos jerárquicos bayesianos son una herramienta flexible para cuantificar las relaciones entre los experimentos originales y las réplicas. La distribución predictiva permite cuantificar la probabilidad de que la réplica tenga éxito y abordar con ella el diseño de la misma.

En este trabajo hemos utilizado modelos normales relativamente sencillos, con el fin de que los detalles técnicos no interfirieran con las ideas planteadas sobre el análisis de la réplica. Obviamente, en la práctica surgirán modelos mucho más complejos en los que habrá de recurrirse a otras técnicas de aproximación e incluso a otras soluciones estadísticas para predecir sobre la réplica.

Son muchas las posibilidades de la replicación en la resolución de problemas importantes, y muy diversos los objetivos con que un investigador puede justificar su práctica.

Si bien el punto de partida para la aplicación de nuestro análisis ha sido un supuesto de réplica ofrecido por Tversky y Kahnemann, son muy frecuentes las replicaciones en Sociología, Psicología y Comunicación (véase una amplia recopilación de réplicas en estas disciplinas en Neuliep 1990). Asimismo, en Medicina, Agricultura, etc, son muy comunes las replicaciones con fines diversos: probar una determinada droga/tratamiento sobre poblaciones distintas para comparar comportamientos, ratificar las conclusiones que sobre una droga/tratamiento algún investigador extrajo de un estudio experimental anterior, cuantificar el sesgo que las deficiencias experimentales de un estudio previo introdujeron en las conclusiones inferenciales, etc. La literatura científica está llena de ejemplos concretos de replicaciones.

La utilización estadística eficaz de la réplica y la información contenida en estudios previos similares resultará esencial en la investigación científica. Ese es nuestro

objetivo de cara al futuro: explotar desde el punto de vista estadístico las posibilidades que ofrece la replicación de estudios en el desarrollo de la Ciencia.

APÉNDICE 1

Definido en los modelos M12 y M3 el factor Bayes para resolver un contraste con hipótesis nula puntual (igualdad de poblaciones o existencia de sesgo, respectivamente) por un cociente de densidades normales (verosimilitudes ponderadas),

$$B_{01} = \frac{N(d|0,h)}{N(d|0,h+\eta)},$$

donde:

$* \ \eta = \begin{cases} 2\tau^2 & \text{en M12} \\ \Delta^2 & \text{en M3,} \end{cases}$
$* \ d = x_r - x_o,$
$* \ h = \gamma_o^2 + \gamma_r^2,$

se llega a una simplificación de la forma:

$$B_{01} = \sqrt{1+\frac{\eta}{h}} \cdot \exp\left\{-\frac{d^2 \cdot \eta}{2h(h+\eta)}\right\}.$$

Así, al definir el éxito por conseguir un factor Bayes lo suficientemente grande ($B_{01} > B$) o lo suficientemente pequeño ($B_{01} < 1/B$), la región de éxito para resolver el contraste con un grado de confianza $B(>1)$ vendrá dada por el conjunto de valores de X_r que satisfacen cualquiera de las dos condiciones:

$$(x_r - x_o)^2 \leq -2\,\frac{h(h+\eta)}{\eta} \cdot Log\left\{B\left(1+\frac{\eta}{h}\right)^{-1/2}\right\} \tag{24}$$

$$(x_r - x_o)^2 \geq 2\,\frac{h(h+\eta)}{\eta} \cdot Log\left\{B\left(1+\frac{\eta}{h}\right)^{1/2}\right\}. \tag{25}$$

Teniendo en cuenta los signos de cada uno de los términos de las desigualdades, para que dicha región resulte «**posible**» (una región coherente), será preciso que se verifique la condición:

$$B < \sqrt{1+\frac{\eta}{h}}.$$

Los valores iniciales para las varianzas muestrales (γ_o^2 y γ_r^2) impondrán las restricciones sobre la coherencia de las regiones de éxito definidas en función del «grado de confianza» (B) que haya fijado el experimentador para resolver el contraste.

En nuestro ejemplo, las restricciones nos llevan a que para valores grandes de B, las regiones a que dan lugar valores pequeños de η y de n_r no resultan coherentes.

APÉNDICE 2

Fijado el tamaño para replicar, n_r, a partir de las simulaciones $\{x_r^i\}_{i=1}^N$ de la distribución predictiva para el resultado de la réplica, se pretende generar una muestra (de idéntico tamaño) de simulaciones de la distribución predictiva para la varianza a posteriori $v_\delta^2(x_r, n_r)$ de la diferencia de medias $\delta = \theta_r - \theta_o$.

Dado que la distribución a posteriori para δ se obtiene de:

$$p(\delta|x_o, x_r) = \int p(\delta|x_o, x_r, \tau^2) \cdot p(\tau^2|x_o, x_r)\ d\tau^2, \tag{26}$$

fijados x_r y n_r aproximamos la varianza a posteriori utilizando el método de Monte-Carlo:

A partir de simulaciones $\{\delta_j\}_{j=1}^M$ de la distribución a posteriori $p(\delta|x_o, x_r)$, estimamos la varianza $v_\delta^2(x_r, n_r)$ mediante:

$$\hat{v_\delta}^2(x_r, n_r) = \frac{1}{M-1} \sum_{j=1}^{M} (\delta_j - \overline{\delta})^2,$$

donde $\overline{\delta} = \frac{1}{N} \sum_{j=1}^{M} \delta_j$ es la media de las simulaciones.

Sustituyendo x_r por cada uno de los valores de las simulaciones $\{x_r^i\}_{i=1}^N$, obtendremos una muestra (aproximada) de la distribución predictiva para la varianza a posteriori de la diferencia de medias.

Las simulaciones $\{\delta_j\}_{j=1}^M$ según $p(\delta|x_o, x_r)$ se generan utilizando el método de composición a partir de simulaciones $\{\tau_j^2\}_{j=1}^M$ de la distribución a posteriori $p(\tau^2|x_o, x_r)$. Dichas simulaciones τ_j^2 se obtienen por el método de aceptación-rechazo, dada la forma de su densidad:

$$p(\tau^2|x_o, x_r^i) \propto m(x_r^i|x_o, \tau^2) \cdot p(\tau^2).$$

Referencias en Tanner, 1992.

Nota. El número de simulaciones que hemos utilizado en todo momento es $N = M = 2000$.

Nota. La implementación y representación se han llevado a cabo utilizando programación en C, el paquete de utilidades Mathematica y S-Plus.

REFERENCIAS

[1] **Mayoral, A.M.** (1995). *La replicación de experimentos. Valoración del éxito.* Tesis de Licenciatura. Universitat de València.

[2] **Neuliep, J.W.** (Ed.) (1990). *Handbook of Replication Research in the Behavioral and Social Sciences.* Select Press.

[3] **Silverman, B.** (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

[4] **Tanner, M.A.** (1992). «Tools for Statistical Inference. Observed Data and Data Augmentation Methods». *Lecture Notes in Statistics*, **67**. Eds. J. Berger *et al*. Springer Berlag, Berlin.

[5] **Utts, Jessica** (1991) «Replication and Meta-Analysis in Parapsychology». *Statistical Science 1991*, **6, 4**, 363–403.

ENGLISH SUMMARY

BAYESIAN DESIGN AND ANALYSIS OF REPLICATIONS IN SCIENTIFIC EXPERIMENTATION

M.J. BAYARRI*
M.A. MARTÍNEZ*
Universitat de València

Replication of experiments is a widespread activity in practice. The motivations for replicating are diverse; usually experimenters wish to verify some previous conclusions, but other motivations are also common, as achieving more precision, validating an assumed model, bias detection, studying the change on inferences when some conditions of the population under study changes, etc. Since, previous to the replication, the results of the original experiment are indeed available, it seems sensible to look for an optimal design of the replication based on all relevant information. We provide the minimal sample size such that there is high probability for the experimenter to be successful in her/his replication. We use Bayesian hierarchical models to quantify the relationship between the original and replicated experiments. The probability of success, key for the design of the replication, is derived from appropriate (Bayesian) predictive distributions.

Keywords: Replicación, modelo jerárquico bayesiano, factor Bayes, simulación, aproximación Monte-Carlo.

⋆This work has been parcially supported by the Consellería de Cultura, Educació i Ciència de la Generalitat Valenciana, research grant GV-1081/93.

*M.J. Bayarri y M.A. Martínez. Dpto. Estadística e I.O. Universitat de València. Dr. Moliner, 50. 46100 BURJASSOT. València.

– Received decembre 1995.

– Accepted june 1996.

Replication of experiments is a most usual practice in scientific experimentation. However, and in spite of its widespread use, ther does not seem to exist a systematization of almost anything related to it. Important questions that usually are not systematically addressed nor answered include, among others, the following: what goals are meant to be achieved? What does a successful replication mean? Should an isolated or a meta-analysis be performed? how should the replications be designed? How should the relations among the original and replicated experiments be incorporated?

The problem has been recognized in some areas and an entire meeting was devoted to it (Neuliep, 1990). However, nor even the exact meaning of a successful replication is free from controversies. In this paper, we systematize some of the goals that seem present when replicating. We also quantify what a «successful replication» might mean, and, based on the information provided by the original experiment, we quantify the probability that the replication will be a success. This is a most natural tool for design. In particular, we provide the needed sample size in the replication so that the probability of success is high enough.

Tversky and Kahnemann (as reported in Utts, 1990) performed an empirical study in which a number of scientists were asked about what a surprising $t-$value would be. The result was somehow paradoxical in that what was perceived as a failure in replicating the significant result could indeed be viewed as providing a stronger evidence against the null. We use simplified version of this experience to exemplify all of the calculations in the paper.

The paper is organized in six sections, of which Section 1 is devoted to an Introduction. In Section 2 we state the problem and the scenarios we shall be treating in the rest of the paper. In particular, we assume that it is wished to replicate an original experiment (ε_o) in which data X_o have been observed. The parameter of interest in ε_o is denoted by θ_o. The replicated experiment, data, and parameter are denoted by ε_r, X_r and θ_r respectively. We only consider in the paper three basic scenarios:

S1: The main goal of ε_r is to reproduce the findings in ε_o. A most usual setup for ε_o is that of a hypothesis testing about θ_o and we assume that a similar testing will be performed in ε_r.

S2: The main goal of ε_r is to study the similarities or dissimilarities of the replicated population when compared with the original one. A useful is that of estimation of θ_o in ε_o the purpose of ε_r being to infer about the discrepancy $\delta = \theta_o - \theta_r$ between the population parameters.

S3: Some bias is suspected to have incurred in the original experiment. A replication is carried out with the main goal of detecting such bias.

Many different goals for replicating as well as methods for the statistical analysis of both experiments other than those treated here can be found in Mayoral (1995).

The flexibility of Bayesian hierarchical models allows for easy modeling of the relationships (exchangeability, dependency on covariates, temporal or spatial relations, etc.) among the original and replicated populations that should be taken into account when replicating. We use a three levels hierarchical model. The first leve models the conditional distributions of the observable variables. The second, models the relationship among the parameters indexing both, original and replicated, models. The third level gives the prior distribution of the unspecified hyperparameters on the second and first levels. Bayesian machinery is used to develop the needed posterior and predictive distributions.

Depending on the goal of replication, we shall have different meanings for what a successful replication is. All of them, however, can be expressed by identifying (different) subsets of the sample space of the replication such that if X_r lies on such a subset the experimenter would consider to have succeded when replicating. The probability of success (in its different meanings) can thus be easily computed from the completely specified predictive distribution for X_r, and a sample size for the replication can then be chosen such that this probability is high enough.

Section 3 is entitled «Reproduction and comparisons» and is devoted to analyzing a simple model to deal with scenarios S1 and S2 above. Specifically, X_o and X_r are assumed to be conditionally independent normals with means θ_o and θ_r respectively. These are taken to be exchangeable, which is modeled as a normal with unknown mean (which is given a non-informative, flat prior on the third level) and known variance, which is then given different values to reflect different opinions about the degree of similarity of both population means. Optimal sample sizes are derived, under several conditions, for scenario S1, both, when the replicated data is going to be analyzed by itself, and when it is going to be combined with the results of the original experiment. When we are in scenario S2, we might wish to compare both populations by simply inferring about δ, the difference of means, in which case the replications would be considered successful if the inferences are precise enough. On the other hand, the goal might actually be to directly test de equality of both means, in which case «success» would mean to solve the testing (in one or another direction) with enough confidence. We compute optimal sample sizes for the different goals, scenarios and prior opinions.

Section 4 is devoted to replication for bias detection. A different model is used in this Section, in which an unknown bias, β, is added to the mean, μ, of the original experiment, while the replication is assumed free from bias. μ is then given an non-informative prior while the bias is assumed normal with mean 0 and different variances, which reflect the possible sizes of the suspected bias. A most usual statistical analysis for bias detection consists in testing for 0 bias, and we provide

optimal sample sizes for the replication such that, with high probability, the testing is expected to be resolved with a small enough Bayes factor against one of the two hypotheses.

Section 5 is basically a robustification of the analysis in Section 3. Here, the variance of the population means is assumed unknown and is given an inverse gamma distribution in the third stage. This basically amounts to asessing a student t distribution to the means instead of the normal one used in Section 3, thus making the model more robust and more sensitive to data. This flexibility does not come without a prize, and the resulting model is notably more cumbersome to analyze than the one in Section 3. The analyses in Section 3 are carried out resorting to Monte-Carlo methods to perform most integrations.

Section 6 generalizes Section 4 by allowing the variance of the bias to be unknown and then be given an inverse gamma distribution on the third stage. Again this is equivalent to assess a t distribution for the bias, which is most appropriate. Optimal sample sizes for the replication are provided assuming different prior distributions.