Bayesian Inference for Psychology. Part I: Theoretical Advantages and Practical Ramifications

Eric-Jan Wagenmakers, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Ravi Selker, Quentin F Gronau, Martin Ŝmíra, Savha Epskamp, Dora Matzke, Jeffrey N. Rouder, Richard D. Morey

2018

Los valores de p abundan demasiado, aparentemente. Una crítica es que a menudo se malinterpretan como probabilidades posteriores bayesianas, así que se puede creer erróneamente que con $p < ,05$ se puede rechazar una hipótesis nula, cuando realmente se trata de la posibilidad de encintrar valores al menos tan extremos como los observados suponiendo que $\mathcal{H}_0$ sea cierta. El salto de ahí a aceptar la hipótesis alternativa es inválido. El valor p no toma en cuenta la plausibilidad previa de $\mathcal{H}_0$, ni que los datos insuales con $\mathcal{H}_0$ también pueden ser inusuales con $\mathcal{H}_1$.

Aún así, usar valores de p es popular debido principalmente a la costumbre, el desinterés, la pereza, la manquéz, y la cobardía de los piscólogos. Sin embargo, hay una "crisis de confianza" en la psicología que está llevando a la caída de p como el tirano dominante y la toma de medidas en su contra.

Entre las alternativas a los valores p se encuentra el uso de intervalos de confianza (que no son bayesianos y, por lo tanto, son estúpidos). Sin embargo, los intervalos de confianza tienen la limitación de que presuponen que el efecto buscado existe, y el problema pasa a ser uno de estimación de parámetros más que de prueba de hipótesis. El problema no son las pruebas de hipótesis, sino los valores de p, por lo que no es necesario abandonar a las primeras.

Se propone usar pruebas de hipótesis utilizando Factores de Bayes. Estos permiten comparar la adecuación predictiva de dos modelos mientras se cuantifica el cambio en las creencias que los datos traen a ambos modelos.

Las Maravillas de la Inferencia Bayesiana

Estimación Bayesiana de Parámetros

Los estadounidenses votan más por presidentes más altos ($r = .39$, $p = .007$). El coeficiente de correlación es la variable de interés. En Bayes, la incertidumbre se debe

traducir en una distribución de probabilidad llamada *prior*, que en este caso sin conocimiento previo sería una distribución uniforme de -1 a $+1$. Después, la distribución de priors es combinada con la información de los datos para resultar en una distribución posterior, que representa la incertidumbre sobre p después de haber visto los datos. Si la distribución posterior se estrecha con respecto a la previa, significa que los datos fueron informativos.

"Las distribuciones posteriores bayesianas son superiores a los intervalos de confianza. Fight me." Un intervalo de confianza $X\,\%$ para un parámetro θ es un intervalo generado por n procedimiento que en un amuestra repetida tiene una probabilidad $X\,\%$ de contener el valor de θ. Así, la confianza en él reside en el uso repetido. En cambio, la confianza en el intervalo creíble bayesiano aplica directamente a la situación presente.

Beneficios de la estimación bayesiana de parámetros

Beneficio 1. La estimación bayesiana puede incorporar el conociemiento previo

Al seleccionar una distribución adecuada de prior, el investigador puede introducir información útil y limitaciones en su estimación. Por ejemplo, en una tarea en la que aún ejecutando aleatoriamente el sujeto tenga $50\,\%$ de probabilidad de acertar, es necesario incorporar ese supuesto en el análisis. Student y Fisher mismos dijeron que sus métodos solo aplican en ausencia de conocimiento previo.

En el ejemplo de los presidentes y la estatura, se puede partir del supuesto de que la correlación será positiva y restringir la distribución del prior de 0 a $+1$.

La metodología clásica es incapaz de integraar el conocimiento previo salvo por el caso más simple de una restricción de orden.

Beneficio 2. La estimación bayesiana puede cuantificar la confianza de que θ yace en un intervalo específico

Con la distribución posterior de θ se puede conocer, por ejemplo, qué tanto más probable es un valor de .2 que uno de .4, dado que solo necesita evaluarse la diferencia en alturas en la distribución para esos valores. Igualmente, la probabilidad de que θ caiga entre dos valores equivale a la masa posterior en ese intervalo.

Los intervalos de confianza clásicos no pueden más que proveer un porcentaje de confianza. Es imposible especificar los límites del intervalo y después preguntar por la probabilidad de que el valor real se encuentre en esos límites. Lo que se hace es definir un porcentaje de confianza, y los datos dictan los límites del intervalo, no se puede hacer lo contrario.

Al usar intervalos de confianza, lo más probale es que el valor de θ estpe cerca del centro del intervalo, y es improbable que esté cerca de sus límites. Del mismo modo, si el valor de hecho cae fuera del intervalo, es probable que esté bastante cerca de éste. La teoría usual de intervalos de confianza no hace esto explícito.

Beneficio 3. La estimación bayesiana condiciona sobre lo que se sabe (i.e. los datos)

La estimación bayesiana condiciona sobre los datos específicos bajo consideración, e ignora otros sets de datos hipotéticos. Los intervalos de confianza clásicos se basan en el

desempeño promedio a través de sets hipotéticos de datos. Sin importar el resultado, nos dicen que podemos tener un $X\,\%$ de confianza en el valor obtenido.

Así, una diferencia crucial es que los procedimientos clásicos son pre-datos, mientras que los bayesianos son post-datos.

El problema fundamental al promediar a través de sets hipotéticos de datos se conoce como el problema de *"recognizable/relevant subsets"*. El problema solo se puede superar al hacer conclusiones condicionales a los datos observados, lo que remueve la base conceptual de los análisis clásicos.

> *Nuestro trabajo no es seguir ciegamente una regla que acierta en el 90 % de los casos, sino sacar las conclusiones con más probabilidades de ser correctas para el caso específico a analizar. La distribución muestral de un estimador no es una medida de su confiabilidad para un caso individual, porque las consideraciones de muestras que no han sido observadas no son relevantes par la forma en que debemos razonar la que sí lo fue. Esto no significa que no haya conexiones entre el caso individual y la ejecución a largo plazo, pues si encontramos el procedimiento que es mejor para casos individuales, éste sería también el mejor a la larga. El problema es que lo inverso no se sostiene. Haber encontrado un procedimiento que tiene el mejor ajuste a la larga no significa que esa regla sea la mejor para un caso particular. Cambiar la confiabilidad en pocos casos por la confiabilidad en muchos mejora el desempeño a largo plazo, pero tiene un gran efecto sobre el desempeño en los casos individuales.*

Beneficio 4. La estimación bayesiana es coherente (i.e., no inconsistente internamente

Todas las afirmaciones inferenciales deben ser mutuamente consistentes, es decir, la inferencia bayesiana no depende de la forma en que un problema se encuadra. La coherencia esta garantizada por las leyes de la probabilidad. Dentro del marco de referencia bayesiano, las violaciones a la lógica y el sentido común no pueden ocurrir. Dada la naturaleza secuencial de su análisis, en la que toda distribución posterior se sigue de una *prior*, la coherencia esta garantizada.

En el marco de referencia clásico, el concepto de coherencia no tiene lugar. Los problemas se hacen manifiestos cuando distintas fuentes de información se deben combinar. Su remedio usual suele ser enfocarse en una sola fuente de información. Esto solo esconde el problema, pues todo set de datos puede dividirse en sub-sets, y el resultado debería ser independiente del orden o método para dividirlo.

Beneficio 5. La estimación bayesiana se extiende de forma natural a modelos complicados

Sin importar la complejidad del modelo, la inferencia bayesiana sostiene un solo estimador: la distribución posterior. Cuando esta no se puede obtener analíticamente, se puede hacer tomando muestras usando algoritmos numéricos como *Markov chain Monte Carlo*. Al incrementar el número de muestras, la precisión puede aumentarse a voluntad.

Prueba de hipótesis bayesiana

En la estimación bayesiana de parámetros, la meta es la distribción posterior. Sin

embargo, este marco de referencia presupone que existe una relación entre variables. En el caso de los presidentes, asumir una distribución uniforme de p de -1 a 1 presupone que p es distinto de cero. En el marco de referencia clásico, las pruebas de hipótsis se suelen hacer estableciendo un intervalo de confianza y rechazando un valor nulo del parámetro si este no cae dentro de la región del intervalo de confianza. Esto es incorrecto en la formulación bayesiana. Por ello, los bayesianos deben ir más allá de la distribución posterior.

Se necesitan comparar dos modelos: una hipótesis nula que dicte la ausencia de efecto ($\mathcal{H}_0 : p = 0$) y una alternativa que dicte su presencia. En estadística bayesiana, esta hipótesis necesita dictarse de forma específica como una distribución. Por ejemplo, $\mathcal{H}_1 : p \sim \text{Uniform}(-1, 1)$, es decir, todo valor de p es igualmente verosímil a priori.

Con las hipótesis $\mathcal{H}_0$ y $\mathcal{H}_1$ compitiendo, el proceso de actualizar sus plausibilidades relativas está descrito por una simplificación de la regla de Bayes:

$$\underbrace{\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})}}_{Posterior\ odds} = \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{Prior\ odds} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{Bayes\ factor\ BF_{10}}$$

Las probabilidades $p(\mathcal{H}_1)/p(\mathcal{H}_0)$ son las plausibilidades relativas de ambos modelos antes de haber visto los datos. Tras observar los datos, la plausibilidad relativa es cuantificada por las porbabilidades del modelo posterior, es decir, $p(\mathcal{H}_1 \mid \text{data})/p(\mathcal{H}_0 \mid \text{data})$. El cambio entre las probabilidades *prior* y posteriores recibe el nombre de Factor de Bayes, es decir, $p(\text{data} \mid \mathcal{H}_1)/p(\text{data} \mid \mathcal{H}_0)$. Dada la naturaleza subjetiva de las probabilidades del modelo *prior*, el énfasis está en el grado en que los datos logran modificar las creencias, es decir, en el factor de Bayes. Cuando el Factor de Bayes es igual a 6.33, es 6.33 veces más probable que los datos se hayan originado de $\mathcal{H}_1$ que de $\mathcal{H}_0$. Un factor de Bayes menor a 1 indica supremacía del modelo $\mathcal{H}_0$. Nótese que en la ecuación, el subíndice de BF_{10} indica que el numerador es $\mathcal{H}_1$, y el denominador es $\mathcal{H}_1$.

Una interpretación alternativa está en términos del desempeño relativo de los modelos. Habiendo dos modelos, $\mathcal{H}_0$ y $\mathcal{H}_1$, y dos observaciones, $y = (y_1, y_2)$, el factor de Bayes BF_{10} está dado por $p(y_1, y_2 \mid \mathcal{H}_1)/p(y_1, y_2 \mid \mathcal{H}_0)$. Ambos modelos hacen una prediccio+on sobre los datos, y el mejor predictor es preferido.

En el marco de referencia bayesiano, la evidencia es esencialmente relativa, por lo que tiene poco sentido hablar de la evidencia de una hipótesis sin mencionar las predicciones de una hipótesis alternativa.

En resumen, los factores de Bayes comparan la adecuación predictiva de dos modelos estadísticos que compiten. Se evalúa la evidencia en una escala continua, y se cuantifica el cmabio en las creencias que cada modelo genera. Esto vuelve al factor de Bayes "la solución Bayesiana estándar a los problemas de prueba de hipótesis y selección de modelos". Se puede considerar al factor de Bayes (o su logaritmo) un termómetro de la intensidad de la evidencia.

Beneficios de la prueba de hipótesis Bayesiana

Beneficio 1. El factor de Bayes cuantifica evidencia que los datos proveen para $\mathcal{H}_0$ vs. $\mathcal{H}_1$

El factor de Bayes es comparativo: pesa el apoyo a un modelo comparado con el otro. El

valor p toma en cuenta solamente el valor de la hipótesis nula; el valor de la hipótesis alternativa permanece sin especificar. Así, los datos que son improbables bajo $\mathcal{H}_0$ llevan al rechazo de es ahipótesis, si considerar que podrían ser igualmente improbables bajo $\mathcal{H}_1$. Sin embargo, se acepta $\mathcal{H}_1$.

Beneficio 2. El factor de Bayes puede cuantificar la evidencia a favor de $\mathcal{H}_0$
El factor de Bayes no le asigna ningún significado especial a las hipótesis. Las compara mecánicamente y manifiesta preferencia por el modelo con las predicciones más adecuadas. Esto es importante para la replicación de la investigación, y más aún para distinguir entre *ausencia de evidencia* y *evidencia de ausencia*. Es decir, El factor de Bayes tiene tres posibilidades al comparar dos modelos: (1) evidencia a favor de $\mathcal{H}_0$, (2) evidencia a favor de $\mathcal{H}_1$, y (3) evidencia que no favorece a $\mathcal{H}_0$ ni a $\mathcal{H}_1$. En este último caso, un ejemplo sería un factor de Bayes $BF_{10} = 1.5$, que indica que los datos so apenas 1.5 veces más probables bajo $\mathcal{H}_0$ que bajo $\mathcal{H}_1$. El valor de p es incapaz de hacer esa distinción, y podría resultar en $p = .20$ ya sea que haya un efecto de $\mathcal{H}_1$ o no.

Beneficio 3. El factor de Bayes permite que se monitoree la evidencia mientras se acumulan los datos
Siguiendo con la analogía del Factor de Bayes como termómetro, se puede leer en cualquier momento. El análisis puede detenerse en cualquier momento y por cualquier razón, como cuando la evidencia es suficiente o se terminan los recursos, lo que permite practicidad y ética. En cambio, con el marco de referencia clásico, uno está obligado a deherirse al plan inicial a cualquier costo. Esta mal visto incluso mirar los datos antes de obtener la muestra completa.

Cabe mencionar que el marco de referencia clásico puede modificarse para incluir pruebas secuenciales. Sin embargo, el marco Bayesiano no requiere de modificación alguna.

Beneficio 4. El Factor de Bayes no depende de planes de muestreo ausentes o desconocidos.
El factor de Bayes no se ve afectado por el plan de muestreo. El principio de verosimilitud indica que los Factores de Bayes se pueden computar e interpretar sin importar si las intenciones con que se recolectaron los datos son ambiguas, desconocidas o ausentes. Esto es especialmente útil cuando los datos son obtenidos de un proceso natural y no hay algo tal como un plan de muestreo o experimento.

En el caso de p, las muestras están definidas en un inicio y no pueden aumentarse si esto no era parte del plan original de recolección de datos. Una (mala) solución es considerar cada muestra aumentada como si fuese fija y parte del plan original. Sin embargo, esto resulta en un error de comparación múltiple: cada nueva prueba tiene una nueva probabilidad distinta de cero de rechazar equivocadamente la hipótesis nula, lo que falla en controlar la tasa de error tipo I.

En cambio, el Factor de Bayes puede interpretarse sin importar que los datos hayan sido generados por procesos del mundo real, y puede actualizarse constantemente de manera indefinida.

Beneficio 5. El factor de Bayes no esta "violentamente sesgado" en contra de $\mathcal{H}_0$

El Factor de Bayes proporciona una medida precisa de la adecuación predictiva relativa. La mala adecuación de $\mathcal{H}_0$ por sí misma no es motivo suficiente para aceptar $\mathcal{H}_1$. Esto contrasta con el marco de referencia clásico, que solo considera qué tan inusuales son los datos bajo $\mathcal{H}_0$. Se ha propuesto que esa podría ser una de las causas de la resistencia a su abandono: hacen tarea fácil de encontrar efectos donde quizá no los haya.

Diez objeciones a la prueba de hipótesis de Factor de Bayes

Objeción 1. La estimación siempre es superior a la prueba

Mucho se ha argumentado acerca de cómo uno debería abandonar las pruebas de hipótesis en favor de la estimación de parámetros. Sin embargo, cada una tiene su lugar.

La estimación de parámetros es apropiada cuando la hipótesis nula no es de interés o cuando trabajo previo ha desacreditado más allá de toda duda la posibilidad de una hipótesis nula.

Otros escenarios de investigación presentan problemas legítimos de prueba. Por ejemplo, uno no se pregunta "Asumiendo que el tratamiento funciona, ¿qué tan fuerte es su efecto?", sino "¿el tratamiento funciona, para empezar?".

No se debe intentar estimar nada hasta no estar seguros de que hay algo por estimar.

En resumen, tanto la prueba de hipótesis como la estimación de parámetros son importantes en momentos diferentes de la investigación, o en contextos de investigación diferentes.

Objeción 2. Las pruebas de hipótesis Bayesianas pueden indicar evidencia por efectos pequeños que son prácticamente insiginficantes

Con muestras grandes, incluso efectos prácticamente insiginifcantes pueden ser categorizados como "significativos" o "fuertemente apoyados por los datos".

Desde una perspectiva Bayesiana, se reconoce la dependencia del contexto. Efectos pequeños en ciertos contextos pueden ser sumamente relevantes. La mejor decisión es aquella con la más alta utilidad esperada, así que se puede integrar en el análisis una capa adicional que considere las utilidades. Una solución alternativa es definir la hipótesis nula no como un punto, sino como un intervalo prácticamente irrelevante alrededor de cero.

Objeción 3. Las pruebas de hipótesis Bayesianas promueven decisiones binarias

El Factor de Bayes evalúa la evidencia que los datos proveen para $\mathcal{H}_0$ contra $\mathcal{H}_1$, por lo tanto, el Factor de Bayes se relaciona con la evidencia y no con las decisiones. Las decisiones requieren consideraciones adicionales acerca de la utilidad de los resultados. En otras palabrasm el Factor de Bayes mide el cambio en las creencias traído por los datos, o el cambio en la adecuación predictiva de dos modelos que compiten; en contraste, las decisiones implican la consideración adicional de acciones y sus consecuencias.

Objeción 4. Las pruebas de hipótesis bayesianas no tienen significado bajo

mala especificación
El Factor de Bayes es una medida relativa. Al indicar evidnecia evidencia abrumadora (*overwhelming*) para $\mathcal{H}_1$ sobre $\mathcal{H}_0$, no implica que $\mathcal{H}_1$ sea un modelo aceptable de los datos, sino solamente que es superior a $\mathcal{H}_0$. Su propio desempeño podría ser muy malo. Así, la crítica de que el factor de Bayes no mide el ajuste absoluto es correcta, pero es pertinente a todo el modelamiento estadístico. Antes de sacar conclusiones es necesario inspeccionar a fondo los datos visualmente, analizar residuos, y confirmar en general que el modelo no está mal especificado.

Objeción 5. Los priors vagos son preferibles sobre los priors informados
Los Factores de Bayes no se pueden utilizar con priors extremadamente vagos. Una comparación razonable entre $\mathcal{H}_0$ y $\mathcal{H}_1$ requoere que ambos modelos estén especificados de manera razonable. Un prior vago como $\mathcal{H}_1 : \delta$ Uniform$(-10^{100}, 10^{100})$ significará que, para todo set razonable de datos, la hipótesis nula será favorecida.

Así, el problema no está con los factores de Bayes, sino con las distribuciones de priors irracionales.

Objeción 6. Los priors por defecto no son lo suficientemente subjetivos
Algunos Bayesianos han propuesto priors *por defecto* que se pueden aplicar sin importar el área de la investigación como referencia. Otros Bayesianos más subjetivos argumentan que siempre debe incluirse el conocimiento subjetivo en los análisis.

La metodología más "objetiva" tiene sus ventajes: requiere menos conocimiento y esfuerzo, y facilita la comunicación. Es difícil que en modelos complicados se alcance un conocimiento subjetivo suficiente para una aplicación práctica.

Por último, los resultados de un análisis más objetivo suelen ser consistentes con los de un análisis subjetivo.

Objeción 7. Los priors subjetivos no son lo suficientemente objetivos
Es una objeción frecuente al razonamiento Bayesiano: los priors son subjetivos, y en ciencia uno debe evitar la subjetividad a toda costa. Esto ignora convenientemente que todo el modelamiento estadístico es subjetivo: la elección del tipo de regresión, por ejemplo, está motivada por conocimiento previo y experiencia.

Cuando se debe elegir entre un modelo razonable y subjetivo, y uno irracional pero objetivo, la elección clara es el primero. Los priors por defecto son una solución intermedia: intentan ser razonables sin requerir completa subjetividad.

Objeción 8. Los priors por defecto tienen prejuicio contra tamaños de efecto pequeños
Ciertamente los análisis clásicos dan más soporte a los efectos pequeños, aunque eso suele deberse más a su sesgo contra la hipótesis nula. Además, para tamaños de muestra grandes, los Factores de Bayes tienen garantía de apoyar fuertemente $\mathcal{H}_0$, incluso con efectos pequeños.

Así, los análisis Bayesianos podrían ser engañosos solamente bajo estas condiciones: tamaño de muestra pequeño, tamaño de efecto pequeño, y una distribución de priors que

presuponga que el tamaño del efecto es grande. Y aun así, el grado en que la evidencia será engañosa es pequeño.

Objeción 9. Incrementar el tamaño de la muestra resuelve todos los problemas estadísticos
Incluso experimentos de alto poder (es decir, con muestras grandes) pueden dar resultados no informativos. El poder es un concepto pre-experimental que se obtiene mediante los sets de datos hipotéticos que se podrían encontrar. En contraste, la evidencia es post-experimental y toma en cuenta solo el set de datos que d hecho fue observado.

Objeción 10. Los procedimientos Bayesianos también se pueden hackear
Se ha argumentado que los Factores de Bayes también son vulnerables a efectos de sesgo como el reporte selectivo, uso ad-hoc de transformaciones, remoción de outliers, etc. Por supuesto, esto es completamente cierto. La inferencia Bayesiana es coherente y óptima, pero no es magia, y no puede proteger de la malicia o el mal entendimiento de la estadística.

Conclusiones
La estadística Bayesiana es el futuro, amigo. Permite incorporar información previa, no depende de la intención que tuviese el muestreo, y se puede usar para cuantificar y monitorear la evidencia a favor tanto de $\mathcal{H}_0$ como de $\mathcal{H}_1$, en completo contraste con los procedimientos clásicos. Aun así, hay resistencia a su adopción por múltiples razones. Para mitigar la dificultad aparente de realizar análisis Bayesianos, los autores han dessarrollado JASP.