Introducción a la Estadística Bayesiana

Trabajo Fin de Grado

Grado de Estadística Aplicada

Curso 2014/2015

Autor: Dailos Castellano Marrero

Tutor: Xavier Bardina i Simorra

Universitat Autònoma de Barcelona

Resumen

El objetivo del presente trabajo es explicar de manera detallada y paso a paso diferentes técnicas utilizadas para estimar estadísticos y realizar contrastes de hipótesis referentes a proporciones y medias. Para ello se utilizarán técnicas propias de la estadística bayesiana, con lo que a priori se explicará las características destacables de esta corriente estadística y sus posible puntos fuertes, para a posteriori desarrollar una serie de situaciones en las que se precisaría de un análisis estadístico, con tal de estimar ciertos estadísticos o contrastar determinadas hipótesis, utilizando en todo caso, ejemplos numéricos que faciliten la comprensión de las explicaciones que a su vez puedan demostrar la idoneidad de la estadística bayesiana en determinadas circunstancias.

Resum

L'objectiu del present treball és explicar de manera detallada i pas a pas diferents tècniques utilitzades per estimar estadístics i realitzar contrastos d'hipòtesis referents a proporcions i mitjanes. Per a això s'utilitzaran tècniques pròpies de l'estadística bayesiana, amb el que a priori s'explicarà les característiques destacables d'aquest corrent estadístic i els seus possibles punts forts, per a posteriori desenvolupar una sèrie de situacions en les quals es precisaria d'una anàlisi estadística, amb tal d'estimar certs estadístics o contrastar determinades hipòtesis, utilitzant en tot cas, exemples numèrics que facilitin la comprensió de les explicacions que al mateix temps puguin demostrar la idoneïtat de l'estadística bayesiana en determinades circumstàncies.

Abstract

The objective of this dissertation is to explain, step by step and in detail, several techniques used to estimate statistics and to perform statistical hypothesis testing of proportions and means. To that end, techniques of Bayesian statistics will be used so we will first explain the main characteristics for this trend and its possible strong points. After this, we will explain several situations that require a Bayesian analysis in order to estimate some statistics or compare some hypothesis. Numerical examples will always be used in order to provide an easy comprehension of the explanations which will prove the suitability of Bayesian statistics for some particular cases.

Índice

1. Conceptos generales

Teorema de Bayes

El teorema de Bayes fue desarrollado por Thomas Bayes en 1763 y con él se expresa la probabilidad condicional de un evento aleatorio A dado otro evento B, mediante la distribución de probabilidad condicional del evento B dado A y la distribución de probabilidad marginal de sólo A.

Dicho de otro modo, sea $\{A_1, A_2, \dots, A_i, \dots, A_n\}$ un conjunto de sucesos mutuamente excluyentes y exhaustivos, y tales que la probabilidad de cada uno de ellos es distinta de cero. Sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B|A_i)$. Entonces, la probabilidad $P(A_i|B)$ viene dada por la expresión:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

donde:

- $P(A_i)$ son las probabilidades a priori.
- $P(B|A_i)$ es la probabilidad de B en la hipótesis A_i.
- $P(A_i|B)$ son las probabilidades a posteriori.

Por tanto, el teorema de Bayes hace uso de probabilidades a priori, que son probabilidades subjetivas, que se desarrollan a continuación, probabilidad de B en la hipótesis A_i, verosimilitud propia de la muestra, y una distribución a posteriori, que se alcanza mediante el producto de las dos anteriores ponderadas según la verosimilitud propia de la muestra.

Además, cabe destacar que, cuando $A_1, A_2, \dots, A_k$ son k sucesos mutuamente excluyentes, uno de los cuales ha de ocurrir necesariamente; entonces, la ley de la probabilidad total establece que:

$$P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

En el caso continuo, sería:

$$P(B) = \int_{\Omega} P(B|x)f(x)$$

Donde $f(x)$ es la función de densidad de una variable aleatoria X evaluada en x, $P(B|x)$ es la probabilidad de B suponiendo que X=x y Ω es el posible espectro de valores continuos que puede tomar X.

Dando lugar a la siguiente modificación de la regla o formula de Bayes:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)} \quad \text{ó} \quad P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\int_{i=1}^{k} P(B|A_i)P(A_i)}$$

Probabilidad subjetiva

La estadística Bayesiana se basa en la interpretación subjetiva de la probabilidad. Para ello utiliza la percepción existente, por parte del investigador, como una variable modificadora (distribución a priori) de los datos muestrales, que dan lugar a una distribución (distribución a posteriori) con la que formular inferencias con respecto al parámetro de interés.

El hecho de amoldar los datos muestrales obtenidos en función del criterio del investigador convierte a la estadística Bayesiana en un instrumento altamente controvertido, dado que esto puede interpretarse, como que la estadística bayesiana manipula los datos muestrales con el fin de demostrar lo que uno quiere en lugar de dejar que los datos, por sí solos, demuestren o no el objeto de estudio.

Sin embargo, la aportación subjetiva del investigador no tiene que ser de por sí negativa o ser considerada manipuladora (en su sentido más peyorativo), ya que esta aportación subjetiva que realiza el investigador puede darse a causa de conocimientos previos adquiridos a través de otros estudios anteriores o por la intuición del profesional, que a diario observa la situación objeto de estudio.

Por ello la probabilidad subjetiva no debe ser interpretada, de por sí, como un instrumento inválido que únicamente pretende manipular el método científico, dado que esta puede aportar beneficios al propio método, además de poder ser contrastado a posteriori, con tal de dar validez a la probabilidad subjetiva utilizada en el proceso.

A su vez, una problemática existente en numerosas ocasiones es que las muestras son muy pequeñas, con lo que no se cumple los requisitos exigibles por el Teorema Central del Límite, que nos indica que si n es suficientemente grande, la variable aleatoria $\bar{X} = \sum_{i=1}^{n} X_i \,/n$ tiene aproximadamente una distribución normal con $\mu_{\bar{X}} = \mu$ y $\sigma_{\bar{X}}^2 = \sigma^2/n$. Al mismo tiempo, no siempre se puede conocer la distribución que sigue la muestra y los experimentos no se pueden repetir. Requisitos que no son exigibles para la estadística Bayesiana, con lo que puede ser una herramienta de gran utilidad, si no única, en ciertas condiciones.

Dicho lo anterior, se puede comenzar a discernir el concepto de distribución a priori, ésta se puede comprender como una distribución que modela los datos muestrales en función de los conocimientos previos existentes, como por ejemplo estudios realizados anteriormente sobre la materia de interés o simplemente, por la intención de aportar cierta información que el investigador considera oportuna.

Por ello, si considera una muestra, que puede ser o no aleatoria $\boldsymbol{X} = (X_1, \dots, X_n)$ con densidad discreta o continua en la familia $f(x, \boldsymbol{\theta})$, con $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k) \epsilon\, \Theta \subset \mathbb{R}^{\mathrm{k}}$. Suponiendo que se tiene información previa sobre $\boldsymbol{\theta}$. Esta información está expresada por medio de una distribución sobre $\boldsymbol{\theta}$ y es esta distribución la que denominamos distribución a priori.

En conclusión, la distribución a priori lo que pretende es aportar información adicional a los datos extraídos de la muestra, de tal forma que complementen la información obtenida de ellos.

Distribución a Posteriori

Por otro lado, la distribución a posteriori $p(\boldsymbol{\theta}|x)$ es, por la ley multiplicativa de la probabilidad, el producto de la función de distribución de probabilidad $p(\boldsymbol{\theta})$ y la función de verosimilitud $p(x|\boldsymbol{\theta})$.

Dicho de otro modo, la probabilidad a posteriori es aquella que resulta de aplicarle conjuntamente la probabilidad a priori (probabilidad subjetiva) y la verosimilitud de los datos (transformación de los datos experimentales en función de la probabilidad subjetiva), entre la probabilidad de los propios datos experimentales.

Intervalo de Probabilidad

En la estadística bayesiana, se conoce por intervalo de probabilidad a algo similar a lo que se conocería como intervalo de confianza en la estadística Frecuentista.

Del lado Frecuentista, el intervalo de confianza hace referencia a probabilidad de que el estimador calculado se encuentre dentro de dos niveles considerados de confianza, donde la confianza dada a este intervalo suele ser, por ejemplo, del 95%. Dicho de otro modo, si repitiéramos el experimento en multitud de ocasiones, el estimador calculado se encontraría dentro del intervalo en el 95% de las ocasiones, mientras que el 5% de las ocasiones estaríamos estimando erróneamente.

El enfoque bayesiano por el contrario es algo distinto, ya que, el método utilizado para su cálculo sería mediante la curva de la función de densidad que se obtiene a posteriori, donde el área bajo dicha curva y entre unos ciertos valores X e Y con cierta probabilidad (por ejemplo, del 95%) constituyen el intervalo de probabilidad del 95%, entre los mencionados puntos (X, Y).

Una vez conocido alguno de los conceptos generales necesarios para poder comprender, de manera básica, la estadística bayesiana. Resulta pertinente pasar a explicar las diferentes técnicas que se pueden aplicar con tal de calcular diferentes estadísticos, según se precise y en función de la necesidad que se pudiera tener en determinados estudios.

Adicionalmente, caber reseñar que en el último capítulo se hará un resumen de las técnicas aplicadas en los temas que se desarrollaran de aquí en adelante, utilizando la herramienta de análisis estadístico R®, en su versión 3.1.2 de 64-bit.

2. Estimación de una proporción (una población)

A modo de ejemplo, supongamos que se desea estimar la proporción de fraudes cometidos por los clientes de una determinada empresa de aguas, donde alguno de sus clientes, con tal de pagar menos en el recibo mensual de agua, manipulan los contadores para hacer que estos contabilicen menos agua de la que realmente se les suministra.

Para asumir dicho ejemplo, vamos a suponer que tras un pequeño estudio llevado a cabo en la ciudad donde la empresa de aguas proporciona el servicio, se ha obtenido los siguientes resultados. De los 1.300.000 puntos de suministro que tiene la empresa, se ha

escogido al azar una muestra de 150 puntos donde realizar una inspección con tal de observar si, en cada uno de ellos, se está cometiendo algún tipo de fraude, observando que en 12 de los cuales, efectivamente se cometía fraude.

Por otro lado, como se ha mencionado anteriormente, en la estadística bayesiana nos encontramos con la necesidad de tener una probabilidad a priori de la proporción θ (conocida por las observaciones o estudios que se han realizado previamente) que nos da la posibilidad de "modificar" o "actualizar" los datos que se han obtenido a partir del estudio realizado, de manera que a la observación realizada, se le pueda incorpora el conocimiento previo existente. De ahí que sea necesaria la aportación de una tasa que sirva de referencia, con lo que para apoyar el estudio se ha solicitado a una asociación de empresas de aguas del país a la que pertenece la empresa, que se facilite la tasa de fraude habitual del país. Dicha tasa de fraude es del 10%.

Como dicha tasa se trata de una aproximación adquirida por los conocimientos de otros estudios realizados por entes del entorno, ésta no carece de cierta incertidumbre, con lo que cabría esperar que la tasa también pudiera ser de 8% ó de un 12%, siendo igual de verosímiles que la proporcionada por la asociación. Mientras que gracias al dato aportado por estos, se sabe que tasas de 50% o más son altamente improbables.

Con lo mencionado hasta el momento ya podemos decir que ya se dispone de todos los datos necesarios para estimar la proporción de fraude de la ciudad, de tal forma que únicamente faltaría desarrollar la distribución a priori a partir de los datos obtenidos y así poder calcular definitivamente la proporción de fraudes que se comete en la ciudad en estudio.

Gracias a los datos recabados, partimos de la base de que la proporción de fraudes puede estar en torno al 10%, con lo que moviéndonos entorno a dicha proporción y en función de la lógica planteada anteriormente con respecto a los demás valores verosímiles, podemos crear la siguiente tabla con los posibles modelos.

	6%	8%	10%	12%	14%
$P(\theta)$	0.1	0.2	0.4	0.2	0.1

Tabla I

De la cual podemos conocer el valor esperado de $\boldsymbol{\theta}$ mediante la siguiente fórmula:

$$P(\theta) = \sum_{i=1}^{k} \theta_i P(\theta_i),$$

Donde k son los posibles valores de θ: $\theta_1, \theta_2, \dots, \theta_k$ y cumplen la condición:

$$\sum_{i=1}^{k} P(\theta_i) = 1$$

Por tanto, el valor esperado de $\boldsymbol{\theta}$ es:

$$\mathrm{E}(\theta) = (0.06)(0.10) + (0.08)(0.20) + (0.10)(0.40) + (0.12)(0.20) + (0.14)(0.10) = 0.10$$

Se puede apreciar, que el valor esperado de la función de probabilidad a priori, refleja claramente la apreciación existente de que la tasa de fraude se encuentra en torno al 10%. Al

mismo tiempo, se observa que los posibles valores dados a las proporciones se sitúan muy cercanos a ese mismo 10%, lo que indica que la dispersión dada a la probabilidad a priori es bastante corta, es decir, se tiene una elevada certidumbre en relación a la tasa que se debería observar. De lo contrario, los valores serían más dispersos, mostrando en tal caso las dudas existentes en torno a la tasa de fraude.

El siguiente paso sería calcular las verosimilitudes, es decir, la probabilidad de que se produzca la observación dada, suponiendo que el modelo expuesto es válido. Para ello, se ha de calcular la función de densidad de una binomial con parámetros n y θ evaluada en x: (0, 1,…,n), utilizando la siguiente fórmula:

$$P(x|n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Donde,

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

De tal forma que se puede aplicar los datos mencionados en el inicio de este ejemplo, véase n=150 y x=12, lo que proporciona una estimación puntual del fraude relativo a la empresa de aguas del 8%.

Por tanto, aplicando para cada una de las suposiciones hechas anteriormente, la fórmula de la densidad de la distribución binomial, podemos obtener la verosimilitud de cada posible modelo. A modo de ejemplo, en el primer supuesto, se haría de la siguiente manera:

$$\binom{150}{12}0.06^{12}(0.94)^{138} = 0.0735$$

De tal forma que si se aplica lo anterior de manera sucesiva, se obtendría la siguiente tabla:

Modelo	$P(\theta)$	Verosimilitud	$P(\theta)x\,Verosimilitud$	P. Posteriori
0.06	0.1	0.0735	0,0074	0,1018
0.08	0.2	0.1192	0,0238	0,3301
0.10	0.4	0.0836	0,0334	0,4630
0.12	0.2	0.0335	0,0067	0,0928
0.14	0.1	0.0089	0,0009	0,0123
Total	1		0.0722	1

Tabla II

A modo de explicación adicional, cabe destacar que para cada uno de los modelos dados, se ha calculado su verosimilitud mediante la función de densidad de la distribución binomial, luego se ha multiplicado la susodicha verosimilitud por la probabilidad dada a priori y tras hacer cada una de las multiplicaciones pertinentes para todos los modelos propuestos, se ha hecho el sumatorio de estos productos ($P(\theta)x\,Verosimilitud$), para por último dividir cada uno de los mencionados productos entre el sumatorio y así obtener la probabilidad a posteriori.

De este modo, se puede apreciar que, como se explicó en el primer apartado (Teorema de Bayes), se puede calcular y por tanto obtener la probabilidad a posteriori a través de la fórmula de Bayes "alternativa", es decir, mediante:

$$P(0.10|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)} =$$

$$= \frac{(0.0836)(0.4)}{(0.0735)(0.1) + (0.1192)(0.2) + (0.0836)(0.4) + (0.0335)(0.2) + (0.0089)(0.1)} =$$

$$=0.463$$

Una vez hecho los cálculos anteriores, se puede recalcular el valor esperado de la tasa de fraude en función de las nuevas probabilidades (a posteriori):

$$\mathrm{E}(\theta) = (0.06)(0.1018) + (0.08)(0.3301) + (0.10)(0.4630) + (0.12)(0.0928) + (0.14)(0.0123) = 0.0917$$

Dicho cálculo da una tasa de fraude esperada del 9.17%, es decir, mayor de la observada (8%), pero a su vez menor de la que se presuponía en un primer instante (10%), con lo que se aprecia claramente la idiosincrasia propia de la estadística bayesiana, y es que ésta no pretende ni quedarse con el único dato proporcionado por el experimento llevado a cabo, ni tampoco quedarse con la intuición o creencia proporcionada por la experiencia, sino más bien intentar aunar los conocimientos empíricos (en su sentido más estricto) y las apreciaciones hechas por los profesionales de la materia en estudio, con el fin de proporcionar la mayor cantidad de datos y experiencias que sea posible, para así intentar lograr una mejor estimación del parámetro que se pretende calcular.

Como alternativa al sencillo supuesto en el que se ha utilizado una cierta distribución discreta para expresar los posibles modelos probabilísticos que pudiera tomar la tasa de fraude θ, y situándonos en una dimensión más realista. Podríamos considerar la posibilidad de que la tasa de fraude se encontrase en un intervalo que fuera de 0% a 100% de manera continua, es decir, en cualquier punto de dicho intervalo real, lo cual es altamente común.

De esta forma nos encontraríamos en un supuesto en el que las probabilidades a priori no se pueden enumerar, dado que estas son infinitas, con lo que el modo de tratarlas sería mediante la correspondiente distribución continua de probabilidad, que pudiera ser cualquiera, siempre y cuando expresara la visión, intuición o experiencia del investigador.

La función de distribución Beta es la más usada en los casos en los que se pretende estimar una proporción. Esta función de distribución depende de dos parámetros, a y b, ambos mayores que 0.

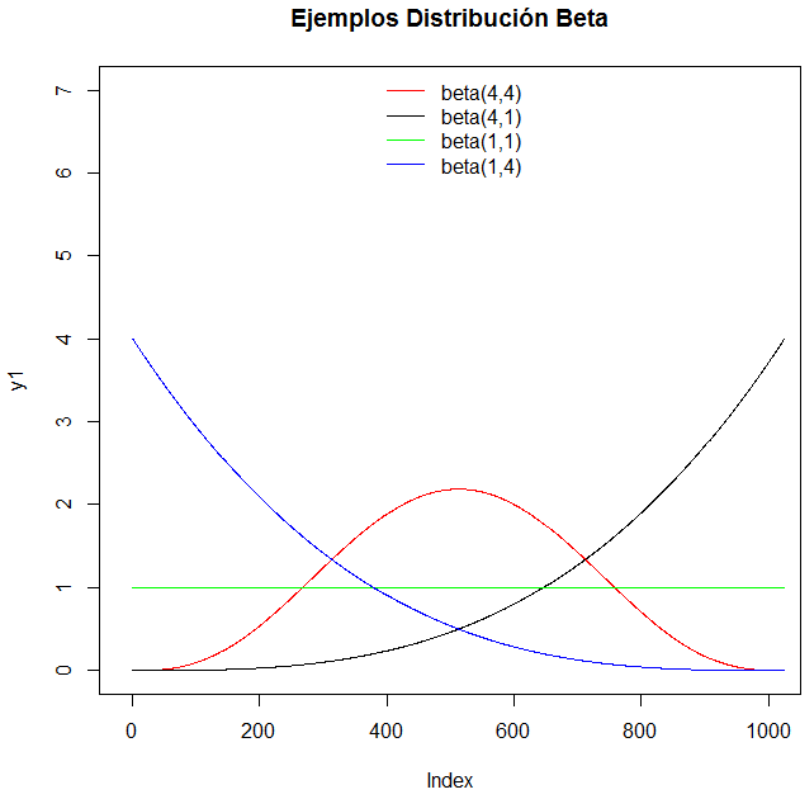
Entre las propiedades de la distribución Beta se encuentran las siguientes:

$$\mu = \frac{a}{a+b} \qquad \sigma^2 = \frac{ab}{(a+b+1)(a+b)^2}$$

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \qquad \text{donde } 0 \leq x \leq 1$$

Adicionalmente, cabe destacar como un caso especial de la distribución Beta aquel en el que ésta toma los valores $a = 1$ y $b = 1$, supuesto en el que coincide con la distribución Uniforme en el intervalo [0, 1].

A modo de ejemplo, se puede apreciar algunas distribuciones Betas dibujadas en la misma gráfica, pero con diferentes parámetros, entre ellas, la equivalente a la distribución Uniforme, es decir, la de color verde, con parámetros Beta(1,1).



Gráfica 1

Como se observa, si el investigador optase por utilizar como distribución a priori la distribución Beta (1, 1), equivalente a una Uniforme en [0, 1], estaría poniendo de manifiesto una posición totalmente neutral, ya que daría el mismo peso a cualquier posible modelo que resultara del análisis, sin ninguna aportación adicional que éste pudiera añadir. En dicho caso, el análisis bayesiano arrojaría el mismo resultado de estimación que el enfoque frecuentista.

Retomando nuestro ejemplo particular, se planteaba que la tasa de fraude esperada 10%, con lo que el siguiente paso (en el caso continuo) sería proponer una distribución Beta que asimilara dicha proporción. Para ello, se utiliza la fórmula de la media de la distribución Beta mencionada anteriormente,

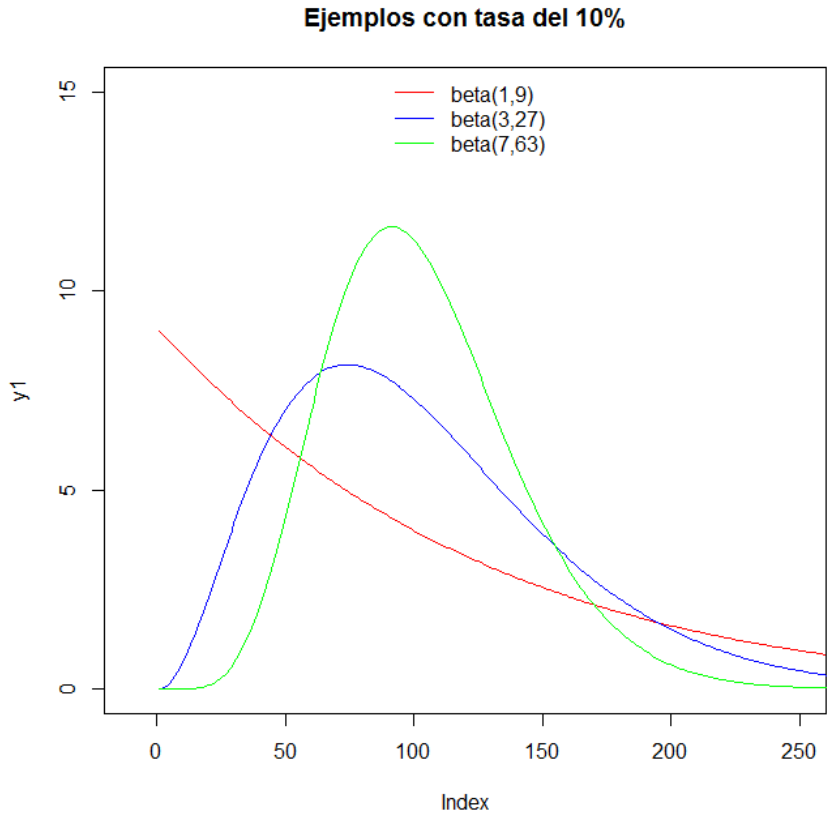
$$\mu = \frac{a}{a+b} = 0.10 = 10\%$$

Como se puede intuir, numerosos valores de a y b podrían dar como resultado la proporción de fraude igual al 10%, con lo que se ha de tomar una decisión. Para ello nos deberíamos basar en la certidumbre que le quisiéramos dar a nuestra probabilidad a priori, ya que a números más bajos de a y b, la distribución a priori presenta más dispersión, con lo que estaría poniendo de manifiesto nuestra incertidumbre.

Como ejemplos, se pude apreciar las siguientes fórmulas y gráficas en las que la media de la distribución Beta toma valor 0.1, es decir, refleja nuestra tasa del 10%, como distribución a priori en función de diferentes grados de certidumbre:

Media (μ):

$$Beta\ (1,9) = \frac{1}{1+9} = 0.10 \qquad Beta\ (3,\ 27) = \frac{3}{3+27} = 0.10 \quad Beta\ (7,\ 63) = \frac{7}{7+63} = 0.10$$



Gráfica 2

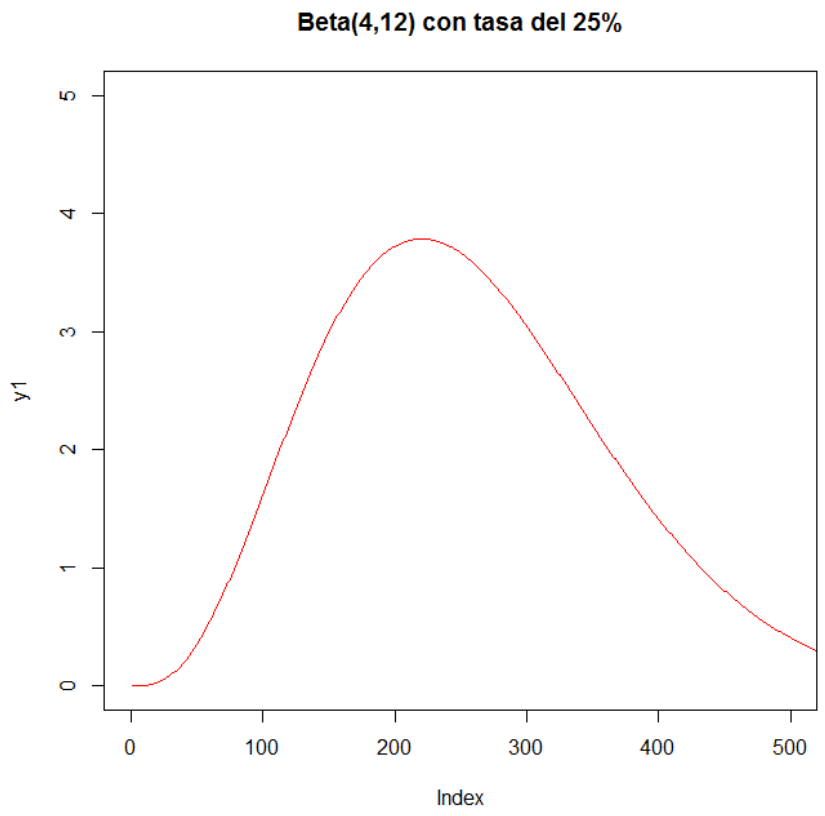
Dado que en nuestro caso nos hemos basado en estudios previos realizados en el mismo país (aunque no en la misma ciudad), se considera que la fiabilidad de nuestra tasa es bastante alta, con lo que se utilizará los parámetros (3, 27).

Por tanto, únicamente nos faltaría mencionar que, debido a que los datos siguen una distribución binomial y la distribución a priori sigue una distribución Beta (a, b), podemos concluir, que la distribución Beta es conjugada para la binomial, transformándose entonces a la distribución a posteriori con los parámetros Beta (a+x, b+(n-x)).

Resumiendo, nos encontraríamos con una probabilidad a priori dada por la distribución Beta (a=3, b=27), con una muestra n=150 y con un total de fraudes detectados entre nuestra muestra de x=12, con lo que conjugadamente la distribución Beta para la binomial de los datos y por tanto, transformando nuestra distribución a priori y obteniendo así, nuestra distribución a posteriori, nos encontraríamos con la distribución Beta (3+12, 27+(150-12)), o lo que es lo mismo, Beta(15, 165), con media igual a 0.0833 y desviación típica igual a 0.00042.

Como se puede observar, nuevamente la distribución a posteriori se acerca notablemente a la tasa de fraude detectada en el estudio realizado, pero como en el caso discreto, desviándose ligeramente hacia un valor algo más cercano a la tasa que a priori se pensaba que podría resultar estimada.

A modo de ejemplo un tanto más extremo, supongamos que la tasa predicha por el investigador fuera del 25% y que la distribución a priori, por tanto, tomara otros parámetros, supongamos Beta (4, 12), que daría lugar a una función de densidad como la siguiente:



Gráfica 3

Bien, si calculáramos nuevamente la distribución a posteriori, ésta nos daría una distribución Beta (16, 150), es decir, una estimación para la tasa de fraude de 0.0934, o lo que es lo mismo, 9.34%.

Como puede apreciarse, a pesar de que la tasa que intuía el investigador es muy dispar en comparación con la recabada por los datos, la determinada por la distribución a posteriori

se acerca mucho más a la muestral debido principalmente a dos causas, una, que la muestra es considerablemente grande, con lo que tiene un peso relativo importante y dos, a que dado que el investigador en este supuesto no se basaba en ningún estudio, coherentemente se decidió poner una distribución a priori con mayor incertidumbre. De lo que se puede concluir, por un lado que actuando de manera coherente (en relación a la distribución a priori) y por otro sabiendo de la importancia del tamaño muestras, los datos corrigen a la probabilidad a priori y que a pesar de que esta tiene cierto peso, la verosimilitud domina en la ponderación total del método bayesiano, con lo que corrige posibles errores de apreciación a priori, aunque no descarta dicha información, sino que la utiliza con el fin de aportar más datos al proceso.

Por otro lado, con tal de ver de manera práctica la importancia del tamaño muestral, utilicemos el mismo ejemplo extremo pero aumentando la muestra, es decir, tomando los siguientes datos: n=1500, x=120, a=4, b=12. Asumiendo dichos datos la tasa de fraude muestral seguiría siendo del 8%, la tasa de la distribución a priori seguiría siendo del 25%, pero la distribución a posteriori cambiaría a Beta (124, 1392) y por tanto la estimación a posteriori del fraude también cambiaría a 0.0818, es decir, a un 8.18%. Demostrando por tanto la importancia del tamaño muestral.

Dicho de otro modo, el peso del tamaño muestral es muy importante en el global del proceso bayesiano, casi igualando el resultado del método al clásico frecuentista, pero a diferencia de éste último método, el método bayesiano es capaz de maximizar su utilidad cuando el tamaño muestral es escaso, siendo por tanto un método ideal en dichos supuestos.

Volviendo al punto en el que nos encontrábamos en el ejemplo inicial, es a partir de este punto en donde podemos sacar conclusiones probabilísticas de nuestro proceso. Expresando directamente en términos de probabilidad, algo imposible en el método frecuentista, las conclusiones. Dicho de otro modo, se puede hablar de que la probabilidad de que $\theta > 7\% = 0.7274$ o que la probabilidad de que $6\% \leq \theta \leq 12\% = 0.8309$. Y es que, al contar con la distribución del parámetro se puede computar el área que se encuentra bajo la función de densidad correspondiente, utilizando para ello un cálculo mediante integrales o mediante software que realicen dichos cálculos de manera más ágil. Por tanto, con la metodología bayesiana se puede calcular la probabilidad de que el parámetro se encuentre en cualquier intervalo dentro del rango [0, 1] (expresado en tanto por uno).

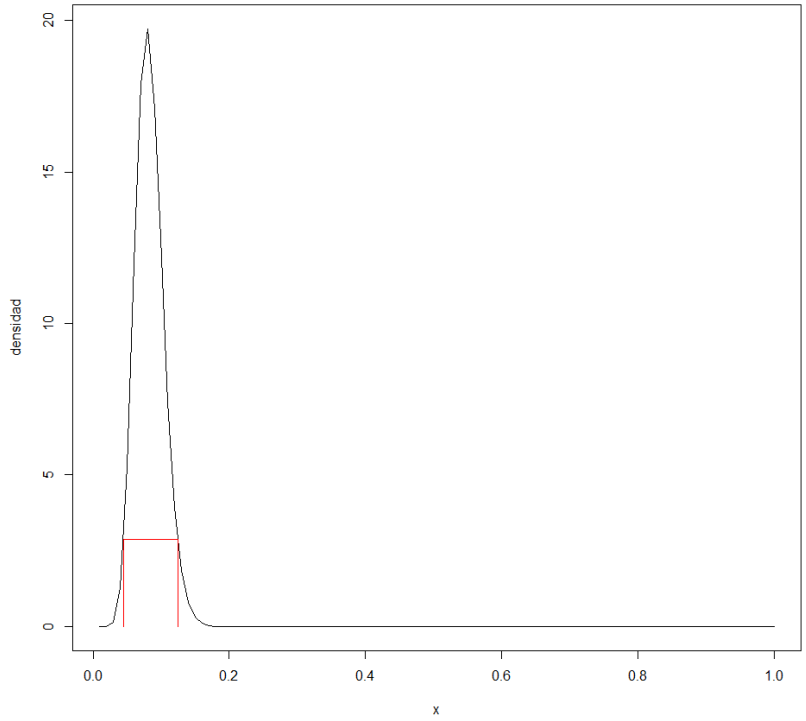
Por último, es especialmente interesante la posibilidad de realizar, en la estadística bayesiana, los intervalos de probabilidad, también llamados intervalos de credibilidad. A diferencia del clásico intervalo de confianza de la estadística frecuentista, el intervalo de probabilidad es un intervalo en el que se encontraría el parámetro que se desea estimar, pero con una cierta probabilidad especificada.

Supongamos que se desea calcular un intervalo que contenga el parámetro θ con probabilidad 1-α, por ejemplo α=0.05, para ello, se calcularía un intervalo donde la probabilidad de que en su interior se encontrara dicho parámetro sería del 95%. El problema que surge es que existirían infinitos intervalos que cumplan dicha condición, con lo que existe un criterio que nos ayuda a decidirnos por ello. Este criterio consiste en escoger aquel intervalo para el cual la función de densidad (en nuestro ejemplo de la distribución Beta(15,165)) cumple la condición de que $f(x) \geq f(y)$ cualquiera sea x perteneciente a dicho

intervalo y cualquiera sea y sin pertenecer a dicho intervalo. De tal modo que este intervalo es el intervalo más corto de entre los que se pueden obtener, cumpliendo la condición de que por ejemplo, bajo el susodicho intervalo se encuentre el parámetro con una probabilidad del 95%.

Dada las infinitas posibilidades de intervalos de credibilidad que se pueden obtener, para calcular el más corto, y por tanto el más denso, es necesario el uso de software especializados, como puede ser el paquete estadístico R®, por tanto, en este caso únicamente se facilitará el intervalo y no se procederá a calcularlo manualmente. Con lo que se tiene un 95% de probabilidad de que el intervalo [0.045, 0.124] contenga el parámetro estimador de la tasa de fraudes.

A continuación se puede ver una apreciación gráfica de dicho intervalo:



Gráfica 4

Cabe destacar nuevamente, que si el tamaño muestral es elevado, dado que la verosimilitud tiene un alto peso, el intervalo de probabilidad no difiere mucho (matemáticamente) del intervalo de confianza, aunque sí conceptualmente. Con lo que una vez más se puede concluir que el método bayesiano es idóneo en el supuesto de muestras pequeñas, al contrario que el método frecuentista.

3. Hipótesis sobre una proporción (una población)

Otra técnica de gran utilidad es aquella que evalúa la hipótesis sobre una proporción. En este caso, se evaluará dicha hipótesis comparando una proporción igual a una constante, pero únicamente en referencia a una población en estudio, es decir,

$$H_0: P = P_0$$
$$H_1: P \neq P_0$$

Para evaluar este contraste de hipótesis en primer lugar se debe conocer ciertos parámetros propios tanto del contraste, como de la estadística bayesiana. Por un lado, sería necesario conocer el valor de P_0 que se querría comprobar en el contraste de hipótesis y por otro, sería necesario conocer la distribución a priori que se le querría imputar a P, que en este caso, como se explico en el tema anterior, se trata de una distribución Beta con parámetros a y b. Cabe destacar, que la media de la distribución Beta (explicada en la fórmula de la página 6) debe dar un resultado aproximado a P_0. Condición razonable, dado que si el investigador pensase a priori que la media o proporción del parámetro fuera distinta de P_0, no tendría sentido hacer el contraste de hipótesis. En resumen, de no cumplirse que:

$$\left|P_0 - \frac{a}{a+b}\right| < 0.015$$

No sería correcto realizar el contraste de hipótesis especificado al inicio de éste capítulo.

Adicionalmente existe otra condición *sine qua non*, y es que los parámetros a y b deben tomar valores mayores que 0, al igual que la media debe comprenderse en el intervalo [0, 1], característico de las proporciones, y la desviación típica ser positiva, siempre y cuando se mantengan los parámetros a y b no negativos.

Como última exigencia, se encuentra la necesidad de establecer una probabilidad a priori de que H_0 se cumpla, que denotaremos por la letra q. Esta probabilidad a priori, al igual que la distribución a priori de P la establece el investigador basándose en los criterios de experiencia y observación, como para la mencionada distribución a priori.

Una vez consideradas las exigencias anteriores, se procedería a obtener los datos propios del experimento y finalmente se calcularía lo que se conoce como el Factor de Bayes a favor de H_0, que lo denotaremos por BF:

$$BF = \frac{{P_0}^x (1-P_0)^{n-x} Beta\ (a,b)}{Beta\ (a+x, b+(n-x))}$$

donde,

$$n = tamaño\ muestral$$
$$x = número\ de\ casos\ favorables$$
$$Beta\ (a,b) = \int_0^1 u^{a-1}\,(1-u)^{b-1} du = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

y donde,

$$\Gamma(x) = \int_0^{\infty} u^{x-1} e^{-u} du$$

Consiguientemente, una vez obtenido el cálculo del Factor de Bayes a favor de H_0, se puede calcular el Factor de Bayes en contra de H_0, que se denotará por BC y se calcula de la siguiente manera:

$$BC = \frac{1}{BF}$$

De tal forma que se puede concluir con el cálculo de la probabilidad a posteriori de la veracidad de H_0. Utilizando el Factor de Bayes a favor de H_0 y la probabilidad a priori de que H_0 es cierta, es decir, q:

$$P(H_0|datos) = \frac{qBF}{qBF + (1-q)}$$

Lo cual se puede transformar a:

$$P(H_0|datos) = \frac{BF}{BF + 1}$$

en el supuesto de que el investigador no tuviera información suficiente para aportar un valor confiable de q y que por lo tanto, decidiera darle a dicho parámetro el valor de $q = 0.5$, lo que equivale a decir que a priori es equiprobable que se dé, tanto la H_0, como la H_1.

Por último, antes de pasar a exponer un caso práctico, convendría mencionar la apreciación de que en la estadística bayesiana, no se pretende escoger entre la hipótesis nula (H_0) y la alternativa (H_1), sino evaluar cuan razonable es escoger una hipótesis frente a la otra, de tal forma que se pueda tomar una decisión en referencia a aquello que se está estudiando o analizando y por tanto actuar en función de dicho análisis.

A modo de ejemplo práctico, consideremos que se desea hacer un estudio de estrés a unos contadores de agua, de una determinada marca y modelo. Para ello, los contadores son sometidos a un estrés mayor al recomendado por el fabricante, es decir, se les hace pasar por ellos un caudal de agua muy elevado (mayor que el máximo recomendado) en un tiempo determinado y se evalúa qué contador sigue funcionando tras la prueba de estrés y qué contador falla y por tanto es inservible.

Para dicho estudio, se desea evaluar el siguiente contraste de hipótesis:

$$H_0: P = 0.75$$
$$H_1: P \neq 0.75$$

es decir, que si la proporción de ruptura es de un 75% o no, que es lo que el fabricante de los contadores establece como tasa de fallo súbito en el supuesto de estudio, considerando además, que dicha H_0 tiene una probabilidad de suceder de q=0.9.

Por otro lado, considerando que el contador se somete a un estrés mayor al recomendado por el fabricante, es lógico pensar que la proporción de rupturas sea mayor a la mitad, pero sin embargo, la experiencia de la empresa suministradora de los contadores dice que estos contadores son más eficientes de lo que el fabricante cree, con lo que nos emplaza a que consideremos como distribución a priori una distribución Beta (7.5, 17.5) que da una media 0.3, lo que es lo mismo, una proporción de rupturas del 30%.

Por tanto, realizamos un estudio sobre 120 contadores, sometiéndolos a un estrés mayor del recomendado por el fabricante y observamos que 54 contadores se han quedado inservibles, lo que da una proporción del 45%.

Una vez obtenidos todos los datos, procedemos a evaluar el contraste de hipótesis y ver si el fabricante se sobreprotege en sus especificaciones, o por el contrario, si el proveedor, nos quiere vender los contadores a pesar de no ser adecuados para el nivel de estrés al que serán sometidos.

Para dicha evaluación, dado que se requiere calcular integrales y cálculos extensos, nos ayudaremos del cálculo computacional, de tal forma que únicamente se expresarán las fórmulas y los resultados finales de dichas formulas. La demostración utilizando el paquete estadístico R se hará en el capítulo de "Estadística Bayesiana con R paso a paso".

En un primer instante sería recomendable hacer un resumen de los datos,

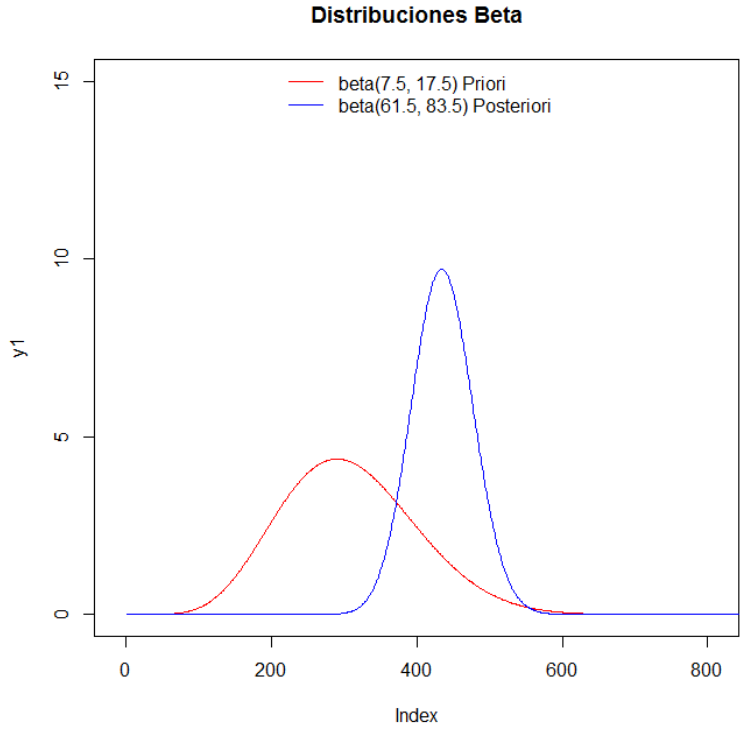
Dato	Valor
P_0	0.75
q	0.9
Beta	(7.5, 17.5)
n	120
x	54

Tabla III

De tal forma que se puede proceder a calcular y dibujar gráficamente las distribuciones Beta a priori y posteriori,

$$Beta\,(7.5, 17.5) = \int_0^1 u^{7.5-1}\,(1-u)^{17.5-1}du = \frac{\Gamma(7.5)\,\Gamma(17.5)}{\Gamma(7.5+17.5)} = 2.5827\ \text{x}\ 10^{-7}$$

$$Beta\,(7.5+54, 17.5+66) = \int_0^1 u^{61.5-1}\,(1-u)^{83.5-1}du = \frac{\Gamma(61.5)\,\Gamma(83.5)}{\Gamma(61.5+83.5)} = 5.053\ \text{x}\ 10^{-44}$$



Gráfica 5

Donde se aprecia que la media o proporción de ambas difieren ligeramente en comparación de la proporción dada por el fabricante. Siendo la proporción de la distribución a priori (proporcionada por el proveedor) del 30% y la posteriori (modificada por los datos) del 42.41%.

A continuación, se calcula el Factor de Bayes a favor de H_0(BF),

$$BF = \frac{0.75^{54}(1-0.75)^{120-54}Beta\ (7.5, 17.5)}{Beta\ (7.5+54, 17.5+(120-54))} = 1.6822 \text{ x } 10^{-10}$$

Con lo que se puede calcular el Factor de Bayes en contra de H_0(BC),

$$BC = \frac{1}{1.6822 \text{ x } 10^{-10}} = 5944541033$$

Para concluir con el cálculo de la Probabilidad a posteriori de la veracidad de H_0,

$$P(H_0|datos) = \frac{0.9x1.6822 \text{ x } 10^{-10}}{0.9x1.6822 \text{ x } 10^{-10} + (1-0.9)} = 1.514\ x\ 10^{-9}$$

Lo que nos arrojaría un resultado bastante favorables hacia el proveedor, ya que la media se acerca a la proporcionada por ellos, más que el resultado del análisis establece que es 5.944.541.033 de veces más probable que sea cierta la hipótesis alternativa, a que lo sea la nula H_0. Además, si se considera la probabilidad a priori que se había dado para este supuesto, se obtiene una probabilidad de $1.514\ x\ 10^{-9}$ de que H_0 sea cierta, es decir, una probabilidad sumamente baja.

Si en lugar de realizar el test de hipótesis para una proporción como valor constante, se decidiera hacer considerando un intervalo, el procedimiento no variaría mucho. Para cerciorarnos de ello supongamos que en el ejemplo anterior el fabricante dijera que la proporción de fallos se encuentra entre [0.45, 0.75] y se mantuviera el resto de variables intactas.

En este caso, el contraste de hipótesis tendría la siguiente estructura:

$$H_0: P \in [P_1 = 0.45, P_2 = 0.75]$$
$$H_1: P \notin [P_1 = 0.45, P_2 = 0.75]$$

Procediendo a evaluar el test de hipótesis anterior, necesitaríamos calcular en un primer instante el área bajo la curva de densidad de la distribución a posteriori que queda a la izquierda de P_1 y P_2, lo que denotaríamos por $F(P_1)$ y $F(P_2)$, lo que requiere de un cálculo computarizado, de tal forma que únicamente se proporcionará el resultado final, como en anteriores ocasiones.

$$F(P_1) = 0.7372 \qquad F(P_2) = 1$$

Una vez obtenido dicho cálculo, se puede obtener la probabilidad de la hipótesis H_0 de la siguiente manera:

$$P(H_0) = F(P_2) - F(P_1) = 1 - 0.7372 = 0.2628$$

Y una vez calculada la probabilidad de la hipótesis H_0, se puede calcular automáticamente la probabilidad de la hipótesis H_1. Sabiendo que el área de la curva de densidad suma (integra) 1 y que el área de $P(H_0)$ es inversa a la de $P(H_1)$, se puede calcular $P(H_1)$ de la siguiente forma:

$$P(H_1) = 1 - P(H_0) = 0.7372$$

Con lo que tras estos pasos es posible realizar el cálculo del Factor de Bayes BF como se expresa a continuación:

$$BF = \frac{P(H_0)}{P(H_1)} = \frac{0.2628}{0.7372} = 0.3565$$

Y en consecuencia la probabilidad a posteriori de la veracidad de H:

$$PP = \frac{qBF}{qBF + (1-q)} = \frac{0.9\ x\ 0.3565}{0.9\ x\ 0.3565 + (1-0.9)} = 0.7624$$

Lo que nos arroja unos resultados, en este caso favorables al fabricante, ya que aproximadamente hay un 65% más de opciones de quedarse con la hipótesis nula que con la alternativa, o dicho desde otra perspectiva, si se considera la probabilidad a priori que se había dado para este supuesto, se obtiene una probabilidad de 0.7624 de que H_0 sea cierta, en este caso, una probabilidad muy alta.

4. Estimación de una diferencia de proporciones (dos poblaciones)

En este capítulo se pretende explicar el procedimiento adecuado para discernir si existe diferencia estadística entre dos proporciones de dos poblaciones distintas mediante técnicas bayesianas.

Al igual que en temas anteriores, sería adecuado plantear un ejemplo numérico con tal de facilitar la comprensión de esta nueva situación.

Supongamos que se desea comparar dos métodos de mantenimiento de contadores, uno que denominaremos por X_1 y que consiste en hacer pasar agua a presión por el susodicho contador, con tal de eliminar impurezas que puedan atascarlo con el tiempo. Y otro método alternativo denominado X_2, que consiste en hacerle pasar unos químicos que eliminen dicha impureza.

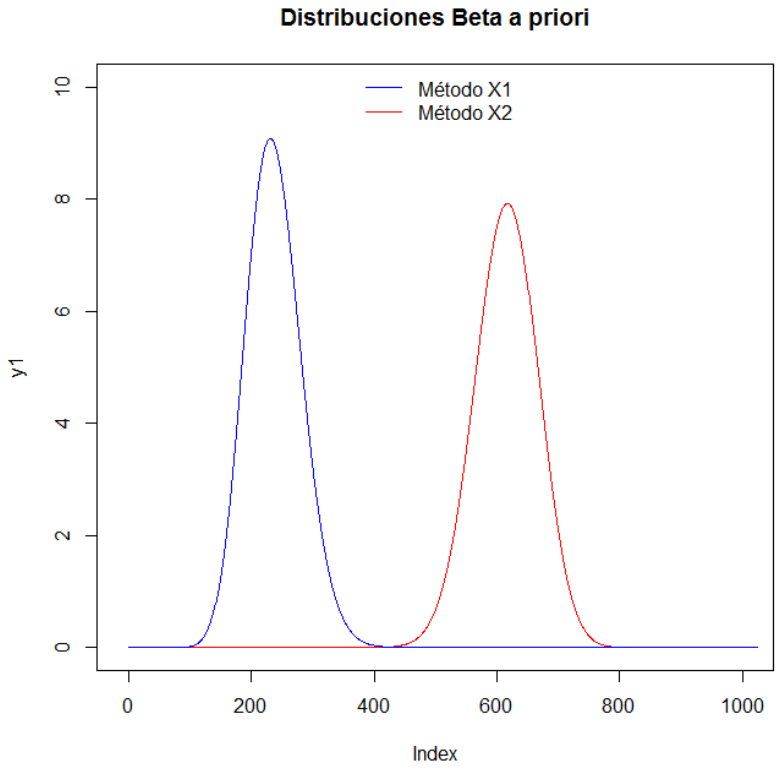
En la actualidad el método utilizado es el X_1 y se conoce que tiene una tasa de efectividad en la limpieza del 30%, es decir, se consigue limpiar significativamente el 30% de los contadores, alargando por tanto la vida útil de estos.

En relación a la probabilidad a priori dada para este ejemplo, cabe mencionar que los investigadores que promueven el estudio creen que el método alternativo proporciona unos resultados mejores, es decir, una mayor tasa de limpieza de los contadores y por tanto alargando la vida útil de más unidades, consiguiendo con ello un ahorro considerable, en renovaciones de contadores, para la empresa.

Por tanto, se tiene una fuerte creencia (basada en la propia observación y experiencia de los operarios de la empresa) de que la proporción del método X_1 se encuentra en el 30% con unos márgenes más o menos definidos en torno al intervalo 10%-50%. Mientras que para el método X_2, los promotores de dicho producto químico indican que la proporción esperada es del 60% con unos márgenes definidos entre el 40% y el 80%. En ambos casos, se considera que existe una probabilidad muy escasa de estar fuera de dichos intervalos.

Al igual que en capítulos anteriores, la distribución de probabilidad empleada para expresar la densidad a priori es la de la distribución Beta, pero con la salvedad de que en este supuesto nos encontramos con dos proporciones y por lo tanto se considerarán dos distribuciones Betas, es decir, una para cada proporción, en función de la creencia del investigador sobre las probabilidades a priori de las proporciones dadas.

Por ello, se considera una distribución a priori Beta (21, 70) para la proporción de X_1 y de Beta (57, 38) para la proporción X_2, que se pueden apreciar en la siguiente gráfica:



Gráfica 6

Para realizar el estudio se ha escogido 300 contadores y se han repartido equitativamente al azar en dos grupos de 150 contadores, a los que se les aplica sendos tratamientos de limpieza, con tal de apreciar cuántos se han limpiado significativamente, arrojando los siguientes resultados:

	Método X_1	Método X_2
n	150	150
x	50	80
n-x	100	70

Tabla IV

Una vez obtenidos los resultados, podemos calcular las probabilidades a posteriori, es decir, la actualización (a partir de los datos) de las probabilidades a priori. Para el método tradicional X_1 resultan ser Beta (21+50, 70+100), y para el método propuesto X_2 son Beta (57+80, 38+70), es decir, Beta(71, 170) y Beta (137, 108) respectivamente. Por tanto, podemos conocer su media y varianza de estas probabilidades a posteriori, que se calculan al igual que en los capítulos anteriores, por tanto:

$$\overline{X_1} = 0.2946 \quad \sigma^2_{X_1} = 0.0009$$

$$\overline{X_2} = 0.5592 \quad \sigma^2_{X_2} = 0.0010$$

Con lo que

$$\overline{X_2} - \overline{X_1} = 0.2646$$

$$\sigma^2_{\overline{X_2}-\overline{X_1}} = 0.0019$$

De tal forma que suponiendo válido el argumento asintótico, resulta que, aproximadamente

$$\overline{X_2} - \overline{X_1} \sim N\,(0.2646, 0.0019)$$

Como, para la $N\,(0,1)$, $P(-1.96, 1.96) = 0.95$, un intervalo de probabilidad 0.95 es

$$\left[0.2646 - 1.96\sqrt{0.0019}, 0.2646 + 1.96\sqrt{0.0019}\right] = [0.1800, 0.3491]$$

Y $0 \notin [0.1800, 0.3491]$, podríamos concluir que existe direrencia probabilística de proporciones en las dos poblaciones y que por tanto, como se aprecia en la diferencia de medias, es más efectivo el tratamiento con químicos, que el tradicional.

Por otro lado, se podría utilizar el método de la simulación, con el que se procederá a generar k observaciones $(p_i^1, p_i^2, \dots, p_i^k)$ de $p_i|x_1, x_2$, para i=1,2 y se aproxima la probabilidad mediante:

$$\frac{cardinal\,\{p_1^1 - p_2^1 \geq c\}}{k}$$

Donde c es una constante que determina un punto mínimo desde el que se puede hacer afirmaciones precisas.

Mediante simulación podemos generar, por ejemplo, 10000 observaciones de p_1 y p_2 $(p_j^i, j \in \{1,2\}, i \in \{1,2,\dots,10000\})$, ordenar los valores $r_i = p_1^i - p_2^i$ y emplear el intervalo $[r_{(25)}, r_{(975)}]$ para dar un intervalo aproximado de probabilidad de 0.95.

Pero dado que dichos cálculos son complejos de llevar a cabo manualmente, se procederá a realizar dichos cálculos en el tema de aplicación en R®, tras explicar todas las técnicas de este trabajo.

5. Hipótesis sobre una diferencia de proporciones (dos poblaciones)

Para desarrollar esta técnica se contrastará la igualdad de proporciones del tipo:

$$H_0: P_1 = P_2$$
$$H_1: P_1 \neq P_2$$

De tal modo que se podrán aplicar técnicas similares a las del capítulo 3 de este trabajo. Y es que, al igual que en dicho capítulo necesitaremos conocer una probabilidad a priori para la validez de la hipótesis nula H_0 (q), para luego definir la distribución Beta (a_i, b_i) a priori para sendas hipótesis, nula y alternativa, donde i hace referencia a cada una de las hipótesis. Por último, se necesitarán los valores relativos al tamaño muestral (n) y al número de veces que se produce el evento de interés(x) para cada una de las proporciones ($P_1\ y\ P_2$).

Por tanto, una vez considerados todos los datos mencionados, se podría calcular el factor de Bayes a favor de H_0. Que como se puede apreciar, difiere del explicado en el nombrado capítulo 3.

$$BF = \frac{Beta(a_0+x_0+x_1, b_0 + (n-x)_0 + (n-x)_1)Beta(a_1, b_1)}{Beta(a_0 + x_0, b_0 + (n-x)_0)Beta(a_1 + x_1, b_1 + (n-x)_1)}$$

Una vez obtenido el factor de Bayes a favor de H_0 (BF), se puede obtener el factor de Bayes en contra de H_0:

$$BC = \frac{1}{BF}$$

Y por último, la probabilidad a posteriori de la veracidad de H_0:

$$P(H_0|datos) = \frac{qBF}{qBF + (1-q)}$$

Con tal de facilitar la comprensión de lo explicado hasta ahora, supongamos el mismo ejemplo que en el capítulo 4, donde se pretendía observar qué método de limpieza de contadores daba mejores resultado en términos proporcionales, pero en este caso, compararemos si es igual un método que el otro o por el contrario, que difieren proporcionalmente. Para ello, asumiremos que la hipótesis nula tiene una probabilidad a priori de validez q=0.3 y que:

$$a_1 = 2 \qquad b_1 = 8 \qquad x_1 = 3 \qquad (n-x)_1 = 12$$

$$a_2 = 4 \qquad b_2 = 6 \qquad x_2 = 6 \qquad (n-x)_2 = 9$$

Con lo que realizando los cálculos mencionados anteriormente se obtendría los siguientes resultados:

$$BF = \frac{Beta(a_1+x_1+x_2, b_1+(n-x)_1 + (n-x)_2)Beta(a_2, b_2)}{Beta(a_1 + x_1, b_1 + (n-x)_1)Beta(a_2 + x_2, b_2 + (n-x)_2)}$$
$$= \frac{Beta(2+3+6, 8+12+9)Beta(4,6)}{Beta(2+3,8+12)Beta(4+6,6+9)} = \frac{Beta(11,29)Beta(4,6)}{Beta(5,20)Beta(10,15)}$$
$$= \frac{5.424\ x\ 10^{-11}\ x\ 0.002}{4.705\ x\ 10^{-6}\ x\ 5.099\ x\ 10^{-8}} = 0.449$$

$$BC = \frac{1}{BF} = \frac{1}{0.449} = 2.229$$

$$P(H_0|datos) = \frac{0.3\ x\ 0.449}{0.3\ x\ 0.449 + (1-0.3)} = 0.161$$

Por tanto, tras actualizar las probabilidades a priori con los datos, se puede apreciar que únicamente existe una probabilidad del 16.1% de que ambos métodos sean iguales y como las proporciones de contadores limpios son del 0.2 y 0.4 para los métodos tradicional y alternativo respectivamente, se puede concluir que el método alternativo es más efectivo que el convencional.

Por otro lado, si en lugar de proporciones puntuales quisiéramos comparar si la diferencia de proporciones se sitúa dentro de un intervalo, la técnica utilizada para realizar dicho análisis sería ligeramente diferente. En este caso utilizaremos el mismo ejemplo y con los mismos datos que los utilizados hasta ahora.

Sin embargo, en este supuesto nos encontraríamos con un contraste de hipótesis algo diferente al planteado anteriormente, siendo este del tipo:

$$H_0: P_2 - P_1 \in [P_3, P_4]$$

$$H_1: P_2 - P_1 \notin [P_3, P_4]$$

Para analizar dicho contraste y determinar en cuál de las dos hipótesis nos encontraríamos, sería adecuado resumir los datos que conocemos e incorporar los que necesitaríamos para este nuevo supuesto.

Para el método de limpieza tradicional se tiene una proporción de contadores limpios del 20%, tras analizar 15 contadores que habían sido sometidos a dicho tratamiento de limpieza y de los cuales se consideraron significativamente limpios 3 y 12 no limpios.

Para el método de limpieza alternativo, se tiene una proporción de contadores limpios del 40%, tras analizar 15 contadores que habían sido sometidos a dicho tratamiento de limpieza y de los cuales se consideraron significativamente limpios 6 y 9 no limpios.

En referencia a la hipótesis nula, se considera que se tiene una probabilidad a priori de validez, para la misma, de q=0.3. Y en relación a las distribuciones a priori, recordar que se le suponían a las distribuciones Beta con los siguientes parámetros:

$$a_1 = 2 \qquad b_1 = 8 \qquad a_2 = 4 \qquad b_2 = 6$$

Una vez recordados los datos existentes, únicamente nos faltaría incorporar un intervalo en el cuál queremos comprobar si se encuentra la diferencia de las proporciones mencionadas, con lo que tras hacer un análisis de costes, con el que podríamos saber si es rentable modificar el sistema de limpieza de contadores, observamos que si la diferencia se encuentra entre un 0 y un 0.15, no sería rentable económicamente realizar el cambio de método, mientras que si se situara fuera de dicho intervalo, sí repercutiría en el aumento de beneficios de la empresa. Por tanto, tenemos $P_3 = 0$ y $P_4 = 0.15$, es decir, el intervalo sería de entre el 0% y el 15%.

Por tanto, una vez recabados todos los datos necesarios, únicamente quedaría evaluar el contraste de hipótesis de la siguiente forma:

Inicialmente, se ha de simular n (por ejemplo 10.000) valores con una distribución a posteriori Beta $\left(a_1 + x_1, b_1 + (n - x_1)\right)$ que denotaremos por $y_{1,i}$ y otros n (los mismos que en la simulación anterior) valores con una distribución a posteriori Beta $\left(a_2 + x_2, b_2 + (n - x_2)\right)$ que denotaremos por $y_{2,i}$. Para realizar esta simulación es necesario la utilización de un paquete estadístico como el R®, dado que de lo contrario sería prácticamente inasumible.

Posteriormente, se calcula las d_i diferencias, es decir $d_i = y_{2,i} - y_{1,i}$, para luego ordenarlas de menor a mayor y de esa forma contabilizar cuantos valores menores que cada una de las P_3 y P_4 existe, de tal forma que si dividimos dicho contador entre 10.000 obtendríamos las $F(P_3)$ y $F(P_4)$, que equivaldría a sendas distribuciones a posteriori evaluadas en P (área bajo la curva de densidad a posteriori que queda a la izquierda de P).

Una vez realizado dichos cálculos con R®, se procedería a evaluar las siguientes ecuaciones, en las que asumiremos las simulaciones hechas con dicho Software:

Probabilidad de la hipótesis nula:

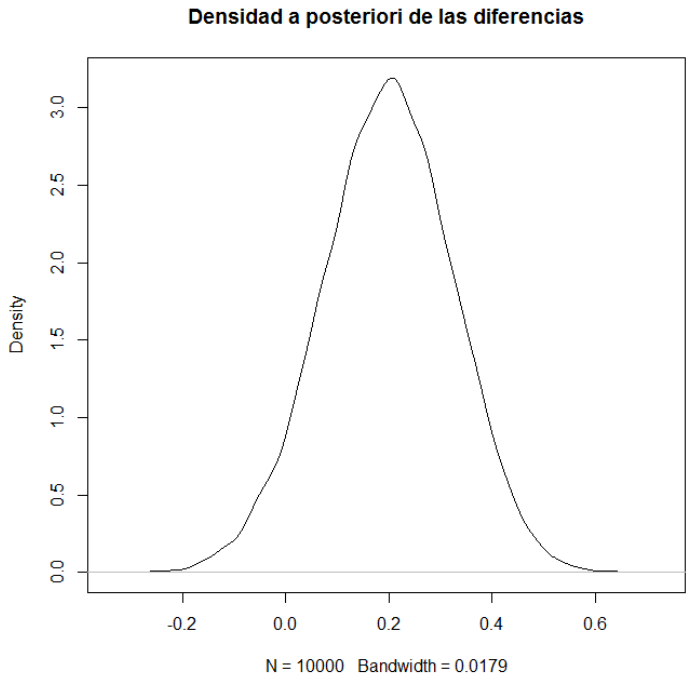
$$P(H_0) = F(P_4) - F(P_3) = 0.2875$$

Factor de Bayes a favor de la hipótesis nula:

$$BF = \frac{P(H_0)}{P(H_1)} = 0.4035$$

Probabilidad a posteriori de la veracidad de la hipótesis nula:

$$PP = \frac{qBF}{qBF + 1 - q} = 0.1474$$



Gráfica 7

Con lo que es 2.48 (1/BF) veces más probable de que sea cierta la hipótesis alternativa a que lo sea la hipótesis nula, es decir, es 2.48 veces más probable que el cambio de método de limpieza sea rentable. Por otro lado, se puede concluir que con una probabilidad a priori como la aportada, se obtiene la probabilidad de un 14.74% de que H_0 sea cierta, o lo que es lo mismo, una probabilidad muy pobre de que la diferencia de media se sitúe en valores no rentables.

6. Estimación de una media (una población)

Habitualmente se necesita calcular medias en lugar de proporciones, ya que la naturaleza propia de los datos no nos lo permite, y es que en ocasiones no se pueda contar el número de veces que se da una cierta condición, sino que se trata de una variable continua que siempre está presente. En estos supuestos, sería ideal estimar una media de la población general a partir de los datos observados en un estudio y si se dispone de alguna información adicional (probabilidad subjetiva), ponerla de manifiesto en la estimación.

Para ilustrar esta técnica, al igual que en las anteriores, se hará uso de un ejemplo numérico, que facilite la comprensión de la misma. Por ello, supongamos que nos encontramos ante una situación en la que una empresa municipal de aguas desea conocer la dureza media del agua que suministra a sus clientes, ya que si se trata de aguas muy duras, estas tienden a realizar deposiciones calcáreas y de magnesio en los contadores, haciendo que estos contabilicen menos agua de la realmente suministrada y se estropeen antes, con el consiguiente gasto derivado a los clientes que supone eso.

En la siguiente tabla se puede ver qué características tienen los diferentes niveles de dureza del agua:

Tipo de Agua	Blanda	Levemente dura	Moderadamente dura	Dura	Muy dura
mg/l	≤ 17	≤ 60	≤ 120	≤ 180	> 180
Característica	Muy Corrosiva	Corrosiva	neutra	Incrustante	Muy Incrustante

Tabla V

En ella se expresa la dureza del agua en mg/l de carbonato cálcico ($CaCO_3$) y el efecto característico que ésta tiene sobre los contadores. De tal forma que un agua muy dura y por tanto muy incrustante es el agua que más deposiciones calcáreas realiza sobre los contadores, haciendo en un principio que estos cuenten más agua de la que realmente pasa por ellos (debido a que estrecha la cavidad por la que pasa y por tanto acelera el flujo), para luego, acortar la vida útil del contador, obligando a su renovación prematura y por tanto aumentando en todo su ciclo de vida el coste para el cliente final. Sin embargo, un agua corrosiva, permite mantener los contadores más limpios, pero deteriora el latón y diversos materiales presentes en la mayoría de instalaciones, además, no tiene altos contenidos de carbonato cálcico y sodio, que hace que el agua potable sea más sana, con lo que tampoco sería lo ideal.

Por tanto, la empresa decide realizar un estudio con tal de conocer el índice medio de dureza en el que se sitúa el agua del municipio. Para ello además, tiene en consideración un estudio rudimentario hecho por unos analistas en prácticas, unas décadas atrás, que arrojó un

valor medio de dureza de 170 mg/l, pero dada las características rudimentarias e inexpertas de dicho estudio, el analistas de la empresa municipal no desean darle una posible veracidad muy elevada, pero mucho menos despreciar la información que proporciona, con lo que se le supone una probabilidad a priori de que la media es mayor que 200 mg/l de un 30%. Adicionalmente, de acuerdo con la tabla de la distribución normal estándar, el valor de z que tiene un 30% de probabilidad a su derecha es 0.52. Así que con los datos mencionados y gracias al método proporcionado por $Berry^1$ se podría calcular la desviación estándar para dicho estudio a través de las siguientes fórmulas:

$$z = \frac{200-170}{s} = 0.52 \qquad \text{y, despejando se obtiene:} \qquad s = \frac{30}{0.52} = 57.69$$

Para finalizar con la recopilación de datos, se decide realizar finalmente el estudio y tras analizar 100 muestras aleatorias de agua, se ha observado que ésta presenta una dureza media de 153 mg/l de $CaCO_3$, con una desviación típica de 20.37 mg/l.

Una vez acabada la recopilación de datos se puede proceder a actualizar la distribución a priori, con los datos muestrales obtenidos y de esa forma obtener la distribución normal actualizada. Puesto que nuestra distribución a priori es normal, se procedería a realizar los siguientes cálculos:

$$c_0 = \frac{1}{s_0^2} \qquad c = \frac{n}{s^2(1+\frac{20}{n^2})^2} \qquad c_1 = c_0 + c \qquad m_p = \frac{c_0 m_0 + c\bar{x}}{c_1} \qquad s_p = \frac{1}{\sqrt{c_1}},$$

donde:

Estadístico	Valor
n	100
$\bar{x}$	153
s_0^2	3328.136
m_0	170

Tabla VI

De tal forma que:

$$c_0 = \frac{1}{s_0^2} = 0.00003 \qquad c = \frac{n}{s^2(1+\frac{20}{n^2})^2} = 0.03 \qquad c_1 = c_0 + c = 0.03003$$

$$m_p = \frac{c_0 m_0 + c\bar{x}}{c_1} = 153.017 \qquad s_p = \frac{1}{\sqrt{c_1}} = 5.77$$

Concluyéndose por tanto que la media de la distribución a posteriori es de 153.017 mg/l de $CaCO_3$ y la desviación típica a posteriori de 5.77, de tal modo que se aprecia una clara influencia de la media muestral debido al tamaño muestral y a la poca credibilidad que se le daba al estudio rudimentario. A su vez, en cuanto a la motivación del estudio, se puede concluir que se trata de un agua dura e incrustante.

En este punto, al igual que en capítulos anteriores, se puede calcular la probabilidad de que la media de mg/l de $CaCO_3$sea menor o mayor que un valor dado, como por ejemplo, la probabilidad de que la media se sitúe en el intervalo [60, 120], es decir, que el agua sea neutra, es del $5.258\ x\ 10^{-9}$ o lo que es lo mismo, una probabilidad casi nula.

7. Hipótesis sobre una media (una población)

En este capítulo se tratará de explicar la realización de un test de hipótesis de la siguiente índole:

$$H_0: m = m_0$$
$$H_1: m \neq m_0$$

Como se puede apreciar, se trata de un test de hipótesis bastante similar al explicado en temas anteriores (3 y 5), con lo que para evitar alargar innecesariamente este trabajo, se procederá a resumir los datos necesarios y los cálculos que se han de realizar, con tal de explicar la técnica apropiada para la realización de éste test de hipótesis y aplicarlos al ejemplo del capítulo 6, pero con la salvedad de que en esta ocasión, en lugar de querer estimar una media de la dureza del agua, querríamos comprobar si esta es igual a 170 o no.

Por tanto, sería necesario recopilar los siguientes datos:

Estadístico	Valor
n	100
$\bar{x}$	153
σ	57.69
m_0	170
$H_0(q)$	0.3
$H_1(t)$	0.7

Tabla VII

Y los cálculos a realizar serían los siguientes:

$$BF = \frac{\frac{\sqrt{n}}{\boldsymbol{\sigma}} \exp\left[-\frac{n}{2\boldsymbol{\sigma^2}}(\bar{x} - m_0)^2\right]}{\left(\frac{\boldsymbol{\sigma}^2}{n} + t^2\right)^{-\frac{1}{2}} exp\left[-\frac{(\bar{x} - m_0)^2}{2(\frac{\boldsymbol{\sigma}^2}{n} + t^2)}\right]} = 0.95$$

$$BC = \frac{1}{BF} = 1.06$$

$$P(H_0|datos) = \frac{qBF}{qBF + (1-q)} = 0.29$$

Es decir, existe una probabilidad del 29% de que la media de dureza del agua se encuentre en 170mg/l de $CaCO_3$.

Si por el contrario, el contraste de hipótesis tratara de dirimir si la media se encuentra en un intervalo o no, en lugar de tomar un valor puntual, el procedimiento cambiaría ligeramente, de tal forma que a diferencia de lo mencionado en el supuesto anterior, en lugar de necesitar una m_0, necesitaremos m_1 y m_2, que podrían tomar los valores 120 y 150 respectivamente, adicionalmente, para este caso se supondrá la inexistencia de información previa válida como una distribución a priori, con lo que se utilizará una distribución Uniforme

U(0, 1), para expresar dicha desinformación. Por tanto, nos quedaríamos con el siguiente test de hipótesis:

$$H_0: m \in [m_1 = 120, m_2 = 150]$$
$$H_1: m \notin [m_1 = 120, m_2 = 150]$$

Y con los siguientes datos:

Estadístico	Valor
n	100
$\bar{x}$	153
σ	57.69
m_1	120
m_2	150
$H_0(q)$	0.3

Tabla VIII

Así pues, para realizar este ejemplo necesitaríamos proceder de manera similar a la utilizada en el capítulo 6 (Estimación de una media), ya que necesitaríamos calcular la media m_p y desviación típica s_p de la distribución a posteriori, mediante la fórmula para el cálculo de c, c_0 y c_1 (descritas en dicho capítulo).

Cabe destacar que c_0 toma valor 0 dado que la distribución a priori sigue una U(0, 1). Con lo que bastaría con calcular c y obtener automáticamente c_1. Pudiendo así conocer los valores que toman m_p y s_p, que serán utilizados para evaluar cada una de las F(m), es decir, la función de distribución evaluada en m, o lo que es lo mismo, el área bajo la curva de densidad a posteriori (con m_p y s_p) que queda a la izquierda de m.

Es recomendable que dichos cálculos se realicen a través de algún Software estadístico, con tal de facilitar el procedimiento. Por ello, a continuación se procederá a resumirlos, aunque podrán verse detalladamente en el anexo de este trabajo, donde se adjuntará la sintaxis (para R®) pertinente para desarrollar cada una de las técnicas aplicadas en los diferentes capítulos.

Por tanto, habría que calcular las siguientes ecuaciones:

Probabilidad de la hipótesis nula:

$$P(H_0) = F(m_2) - F(m_1) = 0.302$$

Factor de Bayes a favor de la hipótesis nula:

$$BF = \frac{P(H_0)}{P(H_1)} = 0.432$$

Probabilidad a posteriori de la veracidad de la hipótesis nula:

$$PP = \frac{qBF}{qBF + 1 - q} = 0.156$$

Con lo que es 2.315 (1/BF) veces más probable de que sea cierta la hipótesis alternativa a que lo sea la hipótesis nula. Por otro lado, se puede concluir que sin tener una probabilidad a priori informativa, se obtiene la probabilidad de un 15.6% de que H_0 sea cierta.

8. Estimación de una diferencia de medias (método exacto - dos poblaciones)

En este capítulo, se explicará cómo estimar una diferencia de medias mediante el método exacto, para dos poblaciones. Por tanto, como se ha realizado anteriormente, se diseñará un ejemplo que ayude a la comprensión de dicha técnica.

Por tanto, a modo de ejemplo, supóngase que en una empresa de aguas, que acaba de ganar un concurso público de concesión para el suministro en una determinada comunidad muy segregada, desea conocer la diferencia de consumos medios de sus futuros clientes en dos explotaciones diferentes de dicha concesión. Para así poder prever la distribución de recursos que necesitará hacer de cara al buen funcionamiento del suministro en ambas explotaciones.

Sin embargo, al ser dos explotaciones nuevas (para la empresa), no se dispone de datos previos, ni de ninguna posible distribución a priori, por lo que para la realización del estudio se asumirá una distribución Uniforme (0, 1), es decir, una distribución no informativa.

En cuanto al diseño del estudio, dado que se trata de una región segregada, para ambas explotaciones se decide utilizar muestras pequeñas con el fin de no encarecer mucho el estudio, con lo que se analiza el consumo histórico facilitados por 25 clientes escogidos al azar en cada una de las dos explotaciones (1 y 2), de las que se desprende los siguientes datos:

Estadístico	Valor
n_1	25
n_2	25
$\bar{x}_1$	35
$\bar{x}_2$	28
s_1^2	0.95
s_2^2	1.05
n	10.000

Tabla IX

Una vez diseñado el estudio y adquiridos los datos, únicamente quedaría desarrollar la técnica que nos incumbe en este capítulo, es decir, estimar la diferencia de medias del consumo de los clientes de las dos explotaciones, mediante el método exacto.

En primer lugar, se calcula $S_1^2 = n_1 s_1^2 = 23.75$ y $S_2^2 = n_2 s_2^2 = 26.25$, para posteriormente generar n=10.000 valores y_{1n}, y_{2n} con distribución χ^2 con $n_1 - 1$ y $n_2 - 1$ grados de libertad respectivamente. Luego se generan dos juegos de n=10.000 valores con distribución Normal estándar, denominados z_1 y z_2, para poder calcular las 10.000 medias simuladas para cada una de las dos explotaciones, con las siguientes fórmulas:

$$m_{1i} = \sqrt{\frac{s_1^2 z_{1i}^2}{y_{1i} n_1}} + \bar{x}_1 \qquad m_{2i} = \sqrt{\frac{s_2^2 z_{2i}^2}{y_{2i} n_2}} + \bar{x}_2$$

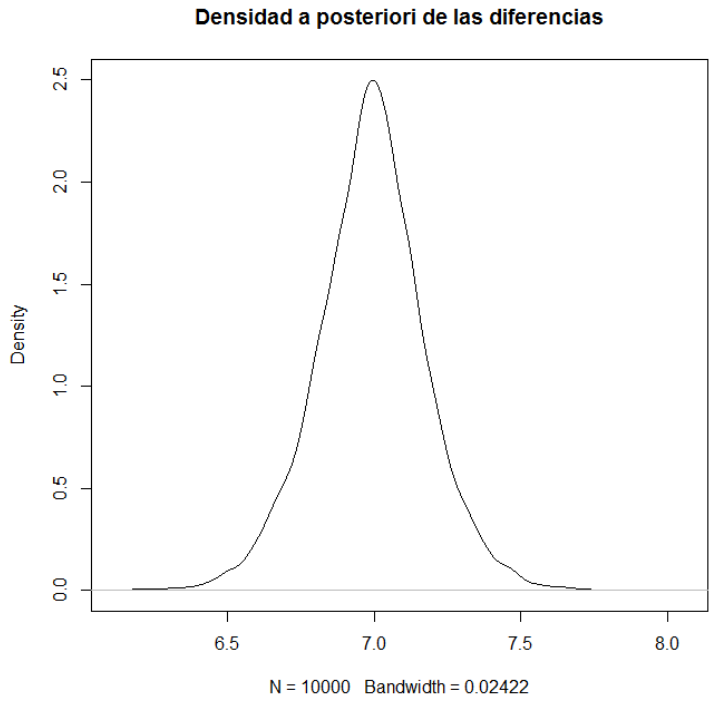
Para finalmente obtener las n diferencias $d_i = m_{1i} - m_{2i}$, de las que si se hace una media, se obtendría el estimador de diferencia de medias.

Dado que dicho proceso es complejo de realizar manualmente, en este supuesto se ha procedido a realizarlo mediante el paquete estadístico R®, con lo que únicamente se mostrará los resultados obtenidos:

Diferencia estimada = 6.99

Percentiles relevantes	
Área	Percentil
2.5	6.611
5	6.678
10	6.760
25	6.878
50	6.993
75	7.105
90	7.217
95	7.296
97.5	7.364

Tabla X



Gráfica 8

Por último, reseñar que la diferencia estimada se ajusta de manera casi exacta a la diferencia que saldría de las dos medias muestrales y que de la distribución empírica a

posteriori de las diferencias es posible obtener estimaciones (no paramétricas) de interés, como pueden ser los percentiles detallados en la tabla anterior.

9. Estimación de una diferencia de medias (método aproximado - dos poblaciones)

Para estimar una diferencia de medias de dos poblaciones mediante el método aproximado, se utilizará el ejemplo explicado en el capitulo anterior, referente al método exacto. Utilizando exactamente los mismos valores muestrales, pero con la salvedad de que en este caso, se planteará tres hipótesis diferentes. Una en la que no se aporte una distribución a priori informativa (como en el caso del capítulo 8), otra en la que se aporta una distribución informativa, mediante la aportación de una media y un desviación típica, para cada una de las poblaciones y por último, otra hipótesis, en la que se aportará una distribución a priori informativa, pero mediante unos determinados valores concretos para cada una de las muestras, con sus respectivas probabilidades, es decir el área que se encuentra a la izquierda de esos puntos.

Para el primero de los casos, dado que la distribución a priori no es informativa, únicamente necesitaremos los datos procedentes de la muestra, es decir:

Estadístico	Valor
n_1	25
n_2	25
$\bar{x}_1$	35
$\bar{x}_2$	28
s_1^2	0.95
s_2^2	1.05
n	10.000

Tabla XI

Una vez recopilados los datos, se procede al cálculo de los estadísticos pertinentes con tal de estimar la diferencia de la media de consumo, en este caso, mediante el método aproximado. Para ello, cabe recordar que al no tener una distribución a priori informativa, únicamente habría que calcular c y no c_0, con lo que c_1 equivaldría a c, que se calcula, para cada una de las poblaciones (explotaciones) de la siguiente forma:

$$c(1) = \frac{n_1}{s_1^2\left(1+\frac{20}{{n_1}^2}\right)^2} = 24.71 \qquad c(2) = \frac{n_2}{s_2^2\left(1+\frac{20}{{n_2}^2}\right)^2} = 22.36$$

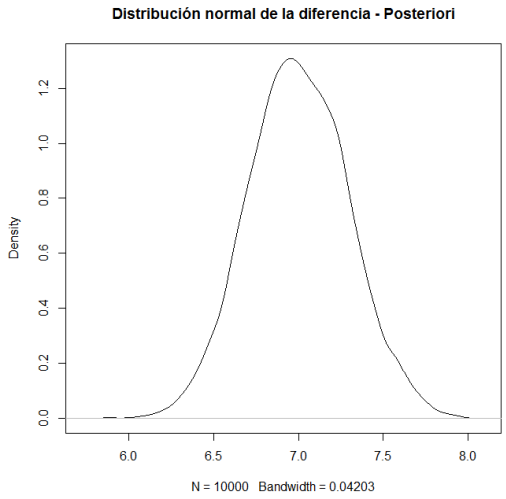
Una vez obtenidos los resultados de las c, se puede proceder a calcular las desviaciones típicas a posteriori de cada una de las poblaciones ($S_{p1}\ y\ S_{p2}$), para luego, calcular la desviación típica a posteriori global, denominada s:

$$S_{p1} = \frac{1}{\sqrt{c(1)}} = 0.20 \qquad S_{p2} = \frac{1}{\sqrt{c(2)}} = 0.21 \qquad s = \sqrt{S_{p1} + S_{p2}} = 0.29$$

Para el cálculo de las medias a posteriori, dado que no hay probabilidad a priori informativa, éstas equivalen a las medias muestrales, de tal forma que únicamente habría que calcular la media conjunta de la diferencia de medias, para así obtener la media a posteriori, es decir:

$$m = m_{p1} - m_{p2} = 7$$

Así pues, se puede concluir que la distribución a posteriori sigue una Normal (7,0.3) y quedaría representada de la siguiente forma:



Gráfica 9

Una vez alcanzado este punto, es posible calcular, tanto el intervalo de credibilidad de por ejemplo $\alpha = 0.05$, que es [6.428, 7.572], o algunos puntos de interés, como pueden ser los percentiles relevantes, que se muestran a continuación:

Percentiles relevantes	
Área	Percentil
2.5	6.428
5	6.520
10	6.626
25	6.803
50	7.000
75	7.197
90	7.374
95	7.480
97.5	7.572

Tabla XII

Si por el contrario, existieran conocimientos previos sobre la materia estudiada y se pudiera aportar cierta información subjetiva, ésta podría también incluirse en el análisis, por ejemplo, conociendo directamente las medias y desviaciones típicas que presentan cada una de las poblaciones en estudio, con lo que a los datos muestrales mencionados anteriormente, habría que añadir dichos parámetros de la distribución a priori, que por ejemplo puede seguir una distribución Normal con:

Estadístico	Valor
n_1	25
n_2	25
$\bar{x}_1$	35
$\bar{x}_2$	28
s_1^2	0.95
s_2^2	1.05
m_1	32
m_2	25
s_1	1
s_2	1.25
n	10.000

Tabla XIII

Dado que en este supuesto sí disponemos de información previa, podemos calcular las c, c_0 y c_1 para cada una de las explotaciones (poblaciones), como se puede apreciar a continuación:

$$c(1) = \frac{n_1}{s_1^2\left(1+\frac{20}{{n_1}^2}\right)^2} = 24.71 \qquad c_0(1) = \frac{1}{s_1} = 1 \qquad c_1(1) = c_0(1) + c(1) = 25.71$$

$$c(2) = \frac{n_2}{s_2^2\left(1+\frac{20}{{n_2}^2}\right)^2} = 22.36 \qquad c_0(2) = \frac{1}{s_2} = 0.64 \qquad c_1(2) = c_0(2) + c(2) = 23.00$$

A su vez, es posible calcular las medias y desviaciones típicas a posteriori, para ambas poblaciones:

$$S_{p1} = \frac{1}{\sqrt{c(1)}} = 0.20 \qquad S_{p2} = \frac{1}{\sqrt{c(2)}} = 0.21 \qquad s = \sqrt{S_{p1} + S_{p2}} = 0.29$$

$$m_{p1} = \frac{c_0(1)m_1 + c\overline{x_1}}{c_1(1)} = 34.88 \qquad m_{p2} = \frac{c_0(2)m_2 + c\overline{x_2}}{c_1(2)} = 28.11 \qquad m = m_{p1} - m_{p2} = 6.77$$

Con lo que una vez llegado a este punto, al igual que en el supuesto anterior podemos calcular, tanto el intervalo de credibilidad de por ejemplo $\alpha = 0.05$, que es [6.209, 7.335], o algunos puntos de interés, como pueden ser los percentiles relevantes, que se muestran a continuación:

Percentiles relevantes	
Área	Percentil
2.5	6.209
5	6.300
10	6.404
25	6.578
50	6.772
75	6.966
90	7.140
95	7.244
97.5	7.335

Tabla XIV

Que como puede apreciarse, si se compara con la tabla de percentiles relevantes del supuesto anterior, los percentiles se han modificado ligeramente a la baja, a causa de la influencia dada por la distribución a priori.

Por último, podría darse el caso en el que la distribución a priori no se proporcionara mediante los valores de las medias y desviaciones típicas de cada una de las poblaciones, sino que esta fuera dada mediante unos valores concretos y unas probabilidades para dichos puntos.

Para poder ilustras este último supuesto, consideraremos los datos anteriores, con la única salvedad de que los datos relativos a las distribuciones a priori son distintos. Se puede apreciar los datos en la siguiente tabla:

Estadístico	Valor	Estadístico	Valor
n_1	25	p_{11}	0.3
n_2	25	x_{12}	34
$\bar{x}_1$	35	p_{12}	0.6
$\bar{x}_2$	28	x_{21}	22
s_1^2	0.95	p_{21}	0.3
s_2^2	1.05	x_{22}	27
x_{11}	32	p_{22}	0.3

Tabla XV

En este caso el proceso variaría ligeramente, ya que hay que calcular previamente las medias y desviaciones típicas a priori. Para ello, en un primer instante hay que calcular la inversa de la distribución normal estándar aplicada a la probabilidad p_i (i=1, 2), es decir, $\psi_i = \phi^{-1}(p_i)$.

$$\psi_{11} = -0.524 \ \psi_{12} = 0.253 \quad \psi_{21} = -0.524 \ \psi_{22} = 0.253$$

Una vez obtenido dichos cálculos, se procede a evaluar las desviaciones típicas y medias a priori de la siguiente forma:

$$s_{01} = \frac{x_{11}-x_{12}}{\psi_{11}-\psi_{12}} = 2.572 \qquad m_{01} = x_{11} - s_{01}\psi_{11} = 32.885$$

$$s_{02} = \frac{x_{21}-x_{22}}{\psi_{21}-\psi_{22}} = 6.429 \qquad m_{02} = x_{21} - s_{02}\psi_{21} = 25.371$$

A partir de este punto, el proceso es idéntico al utilizado en el supuesto anterior, es decir:

$$c(1) = \frac{n_1}{s_1^2\left(1+\frac{20}{n_1{}^2}\right)^2} = 24.71 \qquad c_0(1) = \frac{1}{s_1} = 0.15 \qquad c_1(1) = c_0(1) + c(1) = 24.86$$

$$c(2) = \frac{n_2}{s_2^2\left(1+\frac{20}{n_2{}^2}\right)^2} = 22.36 \qquad c_0(2) = \frac{1}{s_2} = 0.02 \qquad c_1(2) = c_0(2) + c(2) = 22.38$$

$$S_{p1} = \frac{1}{\sqrt{c(1)}} = 0.20 \qquad S_{p2} = \frac{1}{\sqrt{c(2)}} = 0.21 \qquad s = \sqrt{S_{p1} + S_{p2}} = 0.29$$

$$m_{p1} = \frac{c_0(1)m_1 + c\overline{x_1}}{c_1(1)} = 34.98 \qquad m_{p2} = \frac{c_0(2)m_2 + c\overline{x_2}}{c_1(2)} = 28.01 \qquad m = m_{p1} - m_{p2} = 6.98$$

Quedándonos en este supuesto con una media de 6.98 y una desviación típica de 0.29, para la distribución a posteriori Normal. Y con los siguientes percentiles relevantes:

Percentiles relevantes	
Área	Percentil
2.5	6.411
5	6.503
10	6.608
25	6.785
50	6.982
75	7.178
90	7.355
95	7.461
97.5	7.553

Tabla XVI

10. Hipótesis de una diferencia de medias (dos poblaciones)

Para el desarrollo de esta última técnica, se utilizará un ejemplo genérico, en el que se proporcionará únicamente los datos necesarios para su desarrollo. Por tanto, atendiendo al siguiente contraste de hipótesis:

$$H_0: D \in [d_1 = 2.5, d_2 = 5]$$
$$H_1: D \notin [d_1 = 2.5, d_2 = 5]$$

Se recaban los siguientes datos:

Estadístico	Valor
n_1	15
n_2	15
$\bar{x}_1$	18
$\bar{x}_2$	24
s_1^2	2.3
s_2^2	3.4
n	10.000
q	0.4

Tabla XVII

Y se procede a analizarlos de manera similar a la utilizada en los capítulos 7 (Hipótesis sobre una media) y 8 (Estimación de la diferencia de medias por el método exacto), de tal forma que en primer lugar, se calcula $S_1^2 = n_1 s_1^2 = 34.5$ y $S_2^2 = n_2 s_2^2 = 51$, para posteriormente generar n=10.000 valores y_{1n}, y_{2n} con distribución χ^2 con $n_1 - 1$ y $n_2 - 1$ grados de libertad respectivamente. Luego se generan dos juegos de n=10.000 valores con distribución Normal estándar, denominados z_1 y z_2, para poder calcular las 10.000 medias simuladas para cada una de las dos explotaciones, con las siguientes fórmulas:

$$m_{1i} = \sqrt{\frac{S_1^2 z_{1i}^2}{y_{1i} n_1}} + \bar{x}_1 \qquad m_{2i} = \sqrt{\frac{S_2^2 z_{2i}^2}{y_{2i} n_2}} + \bar{x}_2$$

Para finalmente obtener las n diferencias $d_i = m_{1i} - m_{2i}$, de las que si se hace una media, se obtendría el estimador de diferencia de medias (en valor absoluto resulta ser 6.07).

Hasta aquí, el proceso es idéntico al utilizado en el capítulo 8 y a partir de este momento, se asemeja al realizado en el capítulo 7 (con alguna salvedad), es decir, se realizan los siguientes cálculos:

Probabilidad de la hipótesis nula:

$$P(H_0) = F(d_2) - F(d_1) = 0.006$$

Para lo cual se ha tenido que calcular la función de distribución evaluada en d_i, mediante la distribución empírica de las diferencias entre las dos medias.

Factor de Bayes a favor de la hipótesis nula:

$$BF = \frac{P(H_0)}{P(H_1)} = 0.006$$

Probabilidad a posteriori de la veracidad de la hipótesis nula:

$$PP = \frac{qBF}{qBF + 1 - q} = 0.004$$

Dada la complejidad de los cálculos realizados para esta técnica, sería apropiado la utilización de algún Software estadístico que permita ejecutar cálculos masivos, como los empleados. A modo de ejemplo, en el Anexo se adjuntarán la sintaxis utilizada para realizar dichos cálculos, paso a paso.

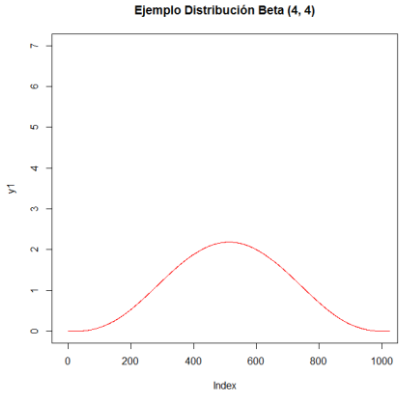
11. Aplicación de las técnicas explicadas en R®

Dada la versatilidad y gratuidad del paquete estadístico R®, éste puede considerarse un Software altamente apropiado para aplicar las diferentes técnicas estudiadas a lo largo de este trabajo. Por ello esa razón, a continuación se hará un resumen de alguna de esas técnicas, explicando la sintaxis que se ha de utilizar para calcular los diferentes estadísticos utilizados en los capítulos anteriores.

Una técnica muy utilizada y de gran utilidad para la estadística bayesiana son las gráficas referentes a la distribución Beta(a, b), distribución muy recurrida para expresar la probabilidad a priori. Esta distribución, se puede expresar gráficamente en R® de la siguiente manera:

```
x <- seq(0, 1, length = 1025) #Proporciona valores al eje X
```

```
y1 <- dbeta(x, 4, 1) #Calcula la densidad de la Beta (4, 1)

plot(y1, col="red", type="l", xlim=c(-10,1010), ylim=c(0,7), main="Ejemplo Distribución Beta (4, 1)")
```

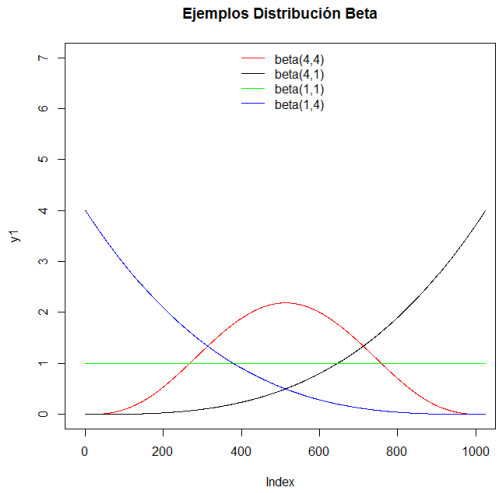


Gráfica 10

Como se puede apreciar, en un primer instante se ha generado un vector con 1025 valores entre 0 y 1 (x), para luego calcular la distribución de densidad de una Beta (4, 1), para cada uno de los valores 'x'. Por último, se ha dado la instrucción pertinente para que dibuje la curva relativa a la distribución de densidad Beta calculada anteriormente, con las indicaciones de que dibujara la línea de color rojo, con una línea continua y que la gráfica mostrara únicamente el rango [-10, 1010], para el eje 'x' y el rango [0, 7], para el eje 'y'. Terminando por incluirle un título a la gráfica.

Si por el contrario se desea expresar varias curvar de densidad de una Beta en una misma gráfica, la sintaxis a utilizar sería la siguiente:

```
x <- seq(0, 1, length = 1025) #Proporciona valores al eje X

y1 <- dbeta(x, 4, 4) #Calcula la densidad de la Beta (4, 4)

y2 <- dbeta(x, 4, 1) #Calcula la densidad de la Beta (4, 1)

y3 <- dbeta(x, 1, 1) #Calcula la densidad de la Beta (1, 1)

y4 <- dbeta(x, 1, 4) #Calcula la densidad de la Beta (1, 4)

plot(y1, col="red", type="l", xlim=c(-10,1010), ylim=c(0,7), main="Ejemplos Distribución Beta")

lines(y2, col="black", type="l") #Añadir la curva de la Beta y2

lines(y3, col="green", type="l") #Añadir la curva de la Beta y3

lines(y4, col="blue", type="l") #Añadir la curva de la Beta y4

legend("top", paste0("beta", c("(4,4)","(4,1)","(1,1)","(1,4)")),

      col=c("red","black","green","blue"), lty=1, bty = "n") #Crear la leyenda explicativa de las curvas
```



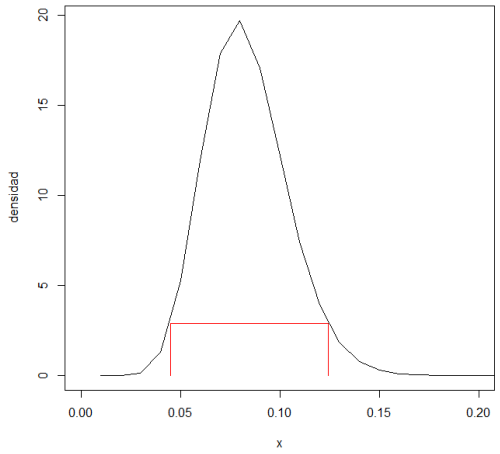
Gráfica 11

En este caso, se ha reproducido lo mencionado para la gráfica anterior, pero con la salvedad de que se ha generado cuatro distribuciones de densidad (y_i) distintas de la distribución de probabilidad Beta y que posteriormente se ha añadido cada una de las y_i a la gráfica con el comando 'lines'. Por último, se ha generado una leyenda que identifique cada una de las y_i.

Otra técnica de gran utilidad y muy utilizada en la estadística bayesiana son los intervalos de credibilidad de mayor densidad posible:

```
a <- 15 #Valor a de la Beta

b <- 165 #Valor b de la Beta

alfa <- 0.05 #Nivel de significación

f <- function(x){

  (dbeta(x[2], a, b) - dbeta(x[1], a, b))^2 +

    (pbeta(x[2], a, b) - pbeta(x[1], a, b) -1 +  alfa)^2

} #Función para desarrollar los intervalos de credibilidad

res <- optim(c(a/(a+b), a/(a+b)), f) #Elección del intevalo de mayor densidad

x <- 1:100 / 100 #Valor dado al eje X

plot(x, dbeta(x, a, b), type = "l", ylab = "densidad") #Gráfica de densidad de la distribución Beta

lines(c(res$par[1], res$par[1]),

      c(0, dbeta(res$par[1], a, b)), col = "red")

lines(c(res$par[2], res$par[2]),

      c(0, dbeta(res$par[2], a, b)), col = "red")

lines(c(res$par[1], res$par[2]),
```

```
    rep(dbeta(res$par[2], a, b), 2), col = "red") # Diferentes líneas que crean gráficamente el intervalo

res # Resultados numéricos del intervalo de probabilidad
```



Gráfica 12

En este caso, se ha empezado por establecer los parámetros de la distribución Beta para la que se desea calcular el intervalo de credibilidad y el nivel de significa al que se desea obtener. A continuación, se ha desarrollado una función que permita calcular el propio intervalo de credibilidad, que será usada en la sentencia siguiente, donde se busca el intervalo de probabilidad de mayor densidad y por tanto el más corto, con un nivel de significación igual al especificado al inicio. Por último, se procede a dibujar la curva de densidad de la distribución Beta, más el marco en el que se encuentra el intervalo de credibilidad (en rojo). Para terminar mostrando los resultados numéricos del intervalo, entre ellos los límites inferior y superior del susodicho.

Por otro lado, en cuanto a las diferentes técnicas explicadas a lo largo de este trabajo, cabe mencionar que éstas han sido calculadas de manera *cuasi* '*manual*', con tal de hacer más comprensible el proceso por el cual se llega a los resultados finales, por lo que se adjuntará la sintaxis que las desarrolla en el anexo final, para así evitar enturbiar este trabajo.

Por último, cabe destacar que en el paquete estadístico R® existen innumerables librerías dedicadas a la estadística bayesiana, que son capaces de ejecutar internamente los cálculos necesarios, evitando de ese modo la necesidad de hacerlo 'manualmente', pero dado que el espíritu de este trabajo ha sido explicar cada una de las técnicas de la manera más detallada posible, no se considera pertinente desarrollar dichas técnicas utilizando librerías que 'automaticen' el proceso.

12. Conclusiones

Como se ha podido apreciar a lo largo de este trabajo, la versatilidad de la estadística bayesiana nos permite poder ampliar los conocimientos adquiridos mediante el muestreo

estadístico, con los posibles conocimientos previos existentes sobre la materia que se estudia, evitando de esa manera que se pierda o desaproveche información existente.

Adicionalmente, la estadística bayesiana tiene una gran ventaja, como alternativa a la estadística frecuentista y es que permite analizar muestras pequeñas sin que esto sea un perjuicio de cara a la estimación "insesgada" de los estadísticos de interés, como se ha podido comprobar a lo largo de este trabajo. Sin embargo, cuando se trata de analizar muestras considerablemente grandes, el método bayesiano no presenta resultados significativamente distintos al método frecuentista debido a que la verosimilitud de los datos muestrales tienen un elevado peso en el global del análisis y por tanto, que en dichos supuestos, sea prácticamente igual considerar un método que el otro.

En otro orden de ideas, el método bayesiano habitualmente conlleva cálculos computacionales complejos, hecho que históricamente ha sido considerado como un hándicap. Sin embargo, en la actualidad, gracias a la existencia de herramientas informáticas avanzadas al alcance de los investigadores, (como puede ser el paquete estadístico R®, explicado en el capítulo anterior) es posible realizar dichos cálculos en un periodo de tiempo breve. Facilitando de ese modo la implantación de las técnicas propias de la estadística bayesiana, donde es habitual realizar múltiples simulaciones, como se ha visto a lo largo de los ejemplos utilizados en este trabajo, donde se llegó a generar en más de una ocasión, para una misma técnica hasta 4 simulaciones de 10.000 valores.

Por otro lado, dado el carácter subjetivo que tiene la distribución a priori, cabe destacar que los estudios llevados a cabo mediante técnicas propias de la estadística bayesiana, deben ser tratados con el mayor rigor científico posible, ya que se trata de una técnica de gran utilidad y por tanto no debe verse empañada por probabilidades subjetivas sesgadas. Aunque, como se mencionó anteriormente, esta posibilidad se ve controlada por la implementación de los datos muestrales que actualizan la susodicha distribución a priori.

Por todo ello, cabe destacar que la existencia y uso de la estadística bayesiana puede ser considerada de gran utilidad, tanto en aquellos supuestos en los que la posibilidad de obtener muestras considerables es casi nula y por tanto no es posible la utilización de la estadística frecuentista con garantías, como en los casos en los que el investigador puede aportar información relevante sobre la materia estudiada, enriqueciendo de esa manera el estudio realizado con dicha información adicional. Pudiendo por tanto coexistir ambos métodos sin que haya ningún impedimento, ya que estos pueden llegar a considerarse complementarios, existiendo la posibilidad de ser usados en partes distintas del proceso investigativo y utilizándose según en qué situaciones uno u otro método.

Bibliografía

- Xunta de Galicia (Consellería de Sanidade) http://www.sergas.es/MostrarContidos_Portais.aspx?IdPaxina=50100
- Dureza del agua: https://es.wikipedia.org/wiki/Agua_dura
- An Introduction to bayesian statistics = Bayesian estatistikarako sarrera = Introducción a la estadística bayesiana. José M. Bernardo Herranz 2004.
- Iniciación a la estadística bayesiana. José Serrano Angulo 2003.
-
- Elementary bayesian biostatistics. Lemuel A. Moyé 2008.
- Berger JO, Berry DA. Statistical analysis and the illusion of objectivity. *American Scientist* 1988.
- Silva LC, Benavides A. El enfoque bayesiano: otra manera de inferir. *Gac Sanit* 2001.
- Benavides A, Silva LC. Contra la sumisión estadística: un apunte sobre las pruebas de significación. *Metas de Enfermería* 2000.
- Bayes T. Essay towards solving a problem in the doctrine of chances. [Reproduced from Phil Trans Roy Soc 1763; 53:370-418]. Studies in the history of probability and statistics. IX. With a bibliographical note by G.A. Barnard. Biometrika 1958.
- Silva LC, Muñoz A. Debate sobre métodos frecuentistas vs bayesianos. Gac Sanit 2000.
- Greenland S. Probability logic and probabilistic induction. Epidemiology 1998.
- Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. Br Med J 1996.

Anexo:

```
## Gráficas ##

x <- seq(0, 1, length = 1025)

y1 <- dbeta(x, 4, 4)

y2 <- dbeta(x, 4, 1)

y3 <- dbeta(x, 1, 1)

y4 <- dbeta(x, 1, 4)

plot(y1, col="red", type="l", xlim=c(-10,1010), ylim=c(0,7), main="Ejemplos Distribución Beta")

lines(y2, col="black", type="l")

lines(y3, col="green", type="l")

lines(y4, col="blue", type="l")

legend("top", paste0("beta", c("(4,4)","(4,1)","(1,1)","(1,4)")),

       col=c("red","black","green","blue"), lty=1, bty = "n")

## Intervalo de Credibilidad ##

a <- 15

b <- 165

alfa <- 0.05

f <- function(x){

  (dbeta(x[2], a, b) - dbeta(x[1], a, b))^2 +

    (pbeta(x[2], a, b) - pbeta(x[1], a, b) -1 +  alfa)^2

}

res <- optim(c(a/(a+b), a/(a+b)), f)

x <- 1:100 / 100

plot(x, dbeta(x, a, b), type = "l", ylab = "densidad")

lines(c(res$par[1], res$par[1]),

     c(0, dbeta(res$par[1], a, b)), col = "red")

lines(c(res$par[2], res$par[2]),
```

```
    c(0, dbeta(res$par[2], a, b)), col = "red")
lines(c(res$par[1], res$par[2]),
    rep(dbeta(res$par[2], a, b), 2), col = "red")
res
## C.H. Prop. Una Población ##
P0=0.7
e=11
n=29
f=n-e
a=3.5
b=1.5
q=0.8
betaAB=(gamma(a)*gamma(b))/(gamma(a+b))
betaAeBf=(gamma(a+e)*gamma(b+f))/(gamma(a+e+b+f))
arriba=(P0^e)*((1-P0)^f)*betaAB
abajo=betaAeBf
BF=(arriba)/(abajo)
BF
BC=1/BF
BC
PdeH=(q*BF)/(q*BF+(1-q))
PdeH
## C.H. Prop. Intervalo de Una Población ##
P1=0.45
P2=0.75
e=54
n=120
```

```
f=n-e

a=7.5

b=17.5

q=0.9

Fp1=pbeta(P1, a+e, b+f)

Fp2=pbeta(P2, a+e, b+f)

PH=Fp2-Fp1

PH

BF=PH/(1-PH)

BF

PP=(q*BF)/(q*BF+1-q)

PP

## Diferencia de Prop. de Dos Poblaciones ##

p1 <- rbeta(10000,71,170)

p2 <- rbeta(10000,137,108)

dif <- p1 - p2

difmas0 <- dif[dif >= 0]

prob <- length(difmas0)/10000

prob

quantile(dif,c(0.025,0.975))

## CH Diferencia de proporciones dentro de un intervalos Dos poblaciones ##

n  = 10000

q  = 0.3

p3 = 0.2

p4 = 0.35

a1 = 72

b1 = 18
```

```
e1 = 30

f1 = 10

a2 = 57

b2 = 38

e2 = 24

f2 = 16

y1 = rbeta(n, a1+e1, b1+f1)

y2 = rbeta(n, a2+e2, b2+f2)

d  = y1-y2

plot(density(d))

d.ord=sort(d, decreasing=FALSE)

plot(density(d.ord))

j=NULL

i=NULL

for (i in 1:10000)

{

        if (d.ord[i] <= p3) j=i

}

j

k=NULL

i=NULL

for (i in 1:10000)

{

        if (d.ord[i] <= p4) k=i

}

k

Fp3 = j/10000
```

```
Fp4 = k/10000

PH  = Fp4-Fp3

PH

BF  = PH/(1-PH)

BF

PP  = (q*BF)/(q*BF+1-q)

PP

## CH de una media dentro de un intervalos Una población ##

m1   = 170

m2   = 178

q    = 0.8

media = 176

var  = 3^2

n    = 10

c  = n/(var*(1+(20/n^2))^2)

mp = media

sp = 1/sqrt(c)

F.m1 = pnorm(m1, mp, sp)

F.m2 = pnorm(m2, mp, sp)

PH = F.m2-F.m1

PH

BF = PH/(1-PH)

BF

PP = (q*BF)/(q*BF+1-q)

PP
```