
# Redes bayesianas aplicadas a la Medicina

**Jose Antonio Férez Rubio**
Master Universitario en Bioinformática y Bioestadística
Área: Bioestadística

**Nombre Consultor/a: Marisa Trullàs Ledesma**

## FICHA DEL TRABAJO FINAL

| | |
|---|---|
| **Título del trabajo:** | *Redes bayesianas aplicadas a la Medicina* |
| **Nombre del autor:** | *Jose Antonio Férez Rubio* |
| **Nombre del consultor/a:** | *Marisa Trullàs Ledesma* |
| **Nombre del PRA:** | *Carles Ventura Royo* |
| **Fecha de entrega (mm/aaaa):** | 06/2017 |
| **Titulación::** | *Master en Bioinformática y Bioestadística* |
| **Área del Trabajo Final:** | *Bioestadística* |
| **Idioma del trabajo:** | *Castellano* |
| **Palabras clave** | *Teorema de Bayes, Redes bayesianas* |


**Resumen del Trabajo (máximo 250 palabras):** *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

El profesor David Sackett definía la Medicina Basada en la evidencia (MBE) como “el uso consciente, explícito y juicioso de la mejor evidencia científica disponible para tomar decisiones sobre los pacientes”. Estas palabras, cargadas de sentido común y espíritu crítico, no eran una simple declaración de buenas intenciones, sino que se convirtieron en la génesis del revolucionario paradigma de la práctica médica en que se ha erigido la MBE. En su propósito por obtener la mejor evidencia posible, el ejercicio de la MBE se ha fundamentado en el método científico, al que considera el mejor “instrumento” para comprender la realidad y expresarla de manera sistemática, inteligible y sintética. Eso sí, sin olvidar ni desdeñar los conocimientos adquiridos por los profesionales de la medicina en su años de ejercicio. De hecho, el principal objetivo de la MBE es integrar la experiencia de los distintos profesionales con la mejor evidencia científica disponible, en aras de mejorar el proceso de toma de decisiones clínicas. Para conseguir este propósito se han creado diferentes entes matemáticos entre los que se encuentran las redes bayesianas, que se han convertido en unas de las “herramientas” más valiosas en el proceso de toma de decisiones.

En este trabajo introduciremos los conceptos y contenidos probabilísticos sobre los que se fundamentan las redes bayesianas, para finalizar con la creación de dos redes asociadas a dos supuestos prácticos.

**Abstract (in English, 250 words or less):**

Professor David Sackett defined Evidence-Based Medicine (EBM) as "the conscious, explicit and judicious use of the best scientific evidence available to make decisions about patients." These words, loaded with common sense and critical spirit, were not simply a declaration of good intentions, but became the genesis of the revolutionary paradigm of medical practice in which MBE was erected. In order to obtain the best possible evidence, the exercise of EBM has been based on the scientific method, which considers the best "instrument" to understand reality and express it in a systematic, intelligible and synthetic way. Of course, without forgetting or neglecting the knowledge acquired by medical professionals in their years of practice. In fact, MBE's main objective is to integrate the experience of the different professionals with the best available scientific evidence, in order to improve the clinical decision-making process. To achieve this purpose, different mathematical entities have been created, including Bayesian networks, which have become one of the most valuable "tools" in the decision-making process.

In this dissertation we introduce the concepts and probabilistic contents on which the Bayesian networks are based, to conclude with the creation of two networks associated to two practical cases.

# Índice

# Índice de figuras



# Índice de tablas

# 1. Introducción

## 1.1 Contexto y justificación del Trabajo

### 1.1.1 Descripción general

Este trabajo pretende ser una guía práctica acerca de la construcción de redes bayesianas en Medicina. Para conseguir este propósito, se ha trazado un itinerario didáctico que permitirá al lector adquirir el bagaje necesario para abordar el problema de la creación de redes bayesianas aplicadas a la toma de decisiones clínicas. Las etapas que componen el itinerario son las siguientes:

- ***El teorema de Bayes y sus derivaciones***: se trata la definición, cálculo, aplicación y derivaciones del teorema de Bayes.

- ***Asociaciones estadísticas y causalidad***: se abordan las distintas asociaciones estadísticas que se pueden hallar entre distintas variables, y se analiza con detenimiento el concepto de causalidad.

- ***Redes bayesianas***: se estudian los conceptos fundamentales asociados a la estructura y "funcionamiento" de las redes bayesianas, para finalizar con la construcción de dos redes bayesianas asociadas a dos supuestos prácticos en Medicina.

### 1.1.2 Justificación del TFM

La toma de decisiones clínicas es un proceso extremadamente complejo, y marcado de manera ineludible por la incertidumbre. Habitualmente, los médicos realizan una análisis de los riesgos y beneficios derivados de cada una de los posibles tratamientos, asociados a los posibles diagnósticos vinculados al caso en cuestión, que está fuertemente condicionado tanto por su experiencia asistencial como por sus conocimientos y creencias. Ante esta perspectiva, se hacía necesario adoptar un nuevo paradigma de razonamiento clínico que abogase por la objetivación en la toma de decisiones y que fuera capaz de incorporar la incertidumbre en sus procesos. Para dar forma a este nuevo modelo, la Medicina recurriría al inestimable potencial de las Matemáticas, representadas principalmente a través de la Estadística y la Probabilidad. En este contexto científico nacieron las redes bayesianas que constituyen una de los "herramientas" matemáticas con mayor potencial en el marco de la toma de decisiones en Medicina.

## 1.2 Objetivos del Trabajo

### 1.2.1 Objetivos generales

Los objetivos generales que persigue este trabajo son los siguientes:

1. Aplicar el teorema de Bayes a situaciones reales de la práctica médica.
2. Identificar las asociaciones estadísticas existentes entre las variables aleatorias manejadas en una determinada situación.
3. Construir redes bayesianas que modelicen problemas o situaciones reales del campo de la Medicina.

### 1.2.2 Objetivos específicos

Cada objetivo general (OG) se concreta a través de una serie de objetivos específicos. A continuación, se detallan los distintos objetivos específicos planteados para el TFM, incluyendo al final de los mismos el objetivo general al que están asociados.

1. Identificar situaciones que sean susceptibles de ser modelizadas con el teorema de Bayes. (OG1)
2. Realizar cálculos probabilísticos que involucren al teorema de Bayes. (OG1)
3. Expresar el teorema de Bayes en sus distintas variantes. (OG1)
4. Diferenciar los conceptos de independencia y correlación. (OG2)
5. Conocer las definiciones de independencia y dependencia condicional. (OG2)
6. Distinguir entre independencia probabilística e independencia causal. (OG2)
7. Definir los fundamentos de las redes bayesianas. (OG3)
8. Asociar modelos canónicos a las relaciones existentes entre variables incluidas en una red bayesiana. (OG3)
9. Implementar redes bayesianas haciendo uso de un software específico. (OG3)

## 1.3 Enfoque y método seguido

La finalidad de este trabajo es glosar el enorme potencial que presentan las redes bayesianas para el proceso de toma de decisiones clínicas. Este propósito se podía haber abordado desde una perspectiva teórica, lo que habría propiciado largas y farragosas demostraciones que habrían desviado la atención de lo verdaderamente importante: la aplicabilidad y utilidad de las redes bayesianas. Por esta circunstancia se ha optado por un enfoque eminentemente práctico, con un marcado carácter didáctico, basado en la claridad, la concisión y la progresividad de los conocimientos tratados, y en contacto permanente con la realidad del lector.

## 1.4 Planificación del Trabajo

### 1.4.1 Tareas

Las tareas que permitirán la consecución de los objetivos generales, y por ende de los objetivos específicos, planteados para el TFM son las siguientes:

- **Documentación y planificación**. Esta tarea comprende todo el proceso de documentación y planificación del TFM, así como de la selección de un software específico para la construcción de redes bayesianas.
- **Teorema de Bayes y sus derivaciones**. Esta tarea abarca todo el proceso de estudio y redacción de los contenidos asociados al teorema de Bayes y sus derivaciones. Esta tarea está vinculada al objetivo general 1 (OG1).
- **Asociaciones estadísticas y causalidad**. Esta tarea abarca todo el proceso de estudio y redacción de los contenidos asociados a las asociaciones estadísticas entre las variables aleatorias y la causalidad. Esta tarea está vinculada al objetivo general 2 (OG2).
- **Redes bayesianas**. Esta tarea comprende tanto el proceso de estudio y redacción de los contenidos asociados a las redes bayesianas, como la modelización de dos supuestos prácticos para su implementación a través de dos redes bayesianas, haciendo uso de un software específico. Esta tarea está vinculada al objetivo general 3 (OG3).
- **Memoria y presentación del TFM**. Esta tarea comprende la revisión y depuración del TFM, así como la preparación de la presentación asociada al mismo.
- **Preparación de la defensa del TFM**. Esta tarea comprende la revisión y depuración de la memoria del TFM y de la presentación asociada en función de la evaluación realizada por el director.

| Tarea | Duración | Fecha inicio | Fecha final |
|---|---|---|---|
| *Documentación y planificación* | 20 días | 24/02/17 | 15/03/17 |
| *Teorema Bayes y sus derivaciones* | 14 días | 16/03/17 | 30/03/17 |
| *Asociaciones estadísticas y causalidad* | 14 días | 31/03/17 | 14/04/17 |
| *Redes bayesianas* | 25 días | 15/04/17 | 09/05/17 |
| *Memoria y presentación del TFM* | 15 días | 10/05/17 | 24/05/17 |
| *Preparación de la defensa del TFM* | 13 días | 25/05/17 | 06/06/17 |

*Tabla 1: Planificación de tareas*

### 1.4.2 Calendario

El siguiente diagrama de Gantt permite visualizar con gran facilidad la planificación de la tareas asociadas al TFM sobre el calendario.

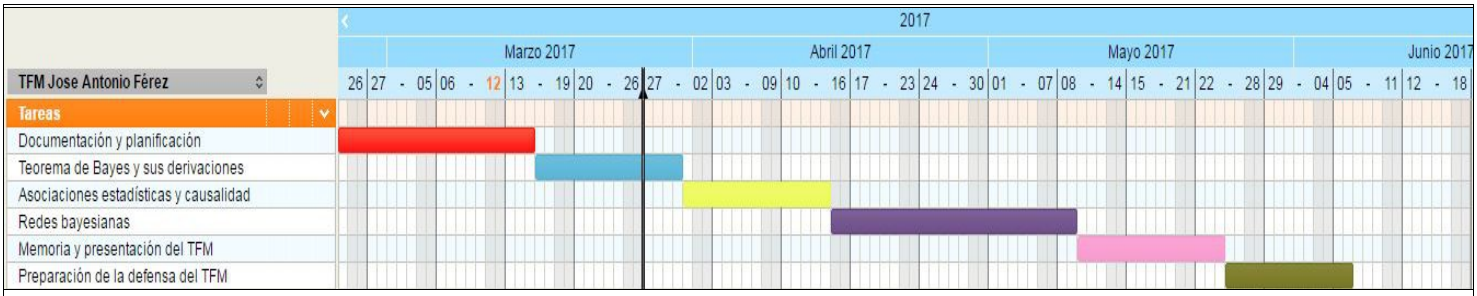



*Figura 1: Diagrama de temporización de tareas*

### 1.4.3 Hitos

Los siguientes hitos sirven de referencia a la hora de valorar el nivel de cumplimiento del plan de trabajo diseñado. Cada uno de los hitos va acompañado de su fecha límite de consecución.

| Hitos | Fecha límite |
|---|---|
| PEC1 *Plan de trabajo* | 15/03/17 |
| PEC2 *Desarrollo de TFM (Fase 1)* | 05/04/17 |
| PEC3 *Desarrollo de TFM (Fase 2)* | 10/05/17 |
| *Memoria y presentación del TFM* | 24/05/17 |
| *Preparación de la defensa* | 06/06/17 |
| *Defensa pública – Tribunal TFM* | 21/06/17 |

*Tabla 2: Secuencia de hitos*

### 1.4.4 Análisis de riesgos

A la hora de planificar y temporizar las distintas tareas asociadas al TFM se han tenido en consideración aquellos factores que pueden repercutir negativamente en el cumplimiento del plan de trabajo. Estos factores, determinantes para el devenir de nuestra planificación, son:

- El enorme volumen de cálculos probabilísticos que conforman el TFM.
- Las dificultades derivadas del manejo del software (suite ofimática, software específico para redes bayesianas, etc.) utilizado para la realización y presentación del TFM.
- Posibles enfermedades leves (de 1 a 3 días)

## 1.5 Resultados esperados

Como resultado del desarrollo del proyecto, se obtendrán al final del mismo los siguientes documentos:

- **Plan de trabajo**: documento que contendrá una descripción general del proyecto, acompañada de una justificación del mismo; los distintos objetivos planteados; el enfoque adoptado; la descripción y planificación de las distintas tareas que propiciarán la consecución de los objetivos propuestos; los hitos establecidos; un análisis de los riesgos que pueden amenazar el cumplimiento de la temporización; y la enumeración de los resultados esperados al final del proyecto.

- **Memoria**: documento que recogerá el trabajo de investigación asociado al TFM. La memoria estará basada tanto en la bibliografía especificada en el apartado del mismo nombre como en el bagaje matemático del autor. Para su realización se utilizará la suite ofimática *OpenOffice*, el software online *draw.io* (https://www.draw.io/), para la realización de grafos y el software *Elvira* (http://leo.ugr.es/elvira/) para la construcción de redes bayesianas.

- **Presentación virtual**: exposición atractiva y didáctica que ilustrará los contenidos y resultados más importantes recogidos en la memoria, de modo que permita mostrar la relevancia científica del TFM. Para la realización de esta presentación se hará uso de software recomendado en el Aula y de la suite ofimática *OpenOffice*.

- **Autoevaluación del proyecto**: documento que contendrá una valoración del cumplimiento del plan de trabajo establecido, de la consecución de los objetivos proyectados, y de las tareas realizadas a lo largo del mismo.

Todos ellos realizados de acuerdo a las indicaciones y plantillas referidas en el Aula.

## 1.6 Estructuración del proyecto

Los estructuración de los capítulos que componen el proyecto se puede visualizar en el índice de este documento.

# 2. Redes bayesianas aplicadas a la Medicina

## 2.1 El teorema de Bayes y sus derivaciones

### 2.1.1 Introducción

Antes de abordar el teorema de Bayes se presentará una situación real que servirá de marco para diferentes ejemplos del apartado.

**Situación real**: en Gabón, la malaria constituye uno de los grandes problemas médicos del país. Esta enfermedad, provocada por un parásito, "determina" una división del país en tres regiones en base al nivel de riesgo de contagio. Además, se ha detectado que los dos grupos sanguíneos asociados a los pacientes que la padecen poseen distintos grados de inmunidad. En base a esta información, se desarrolla un estudio epidemiológico considerando las siguientes variables:

- *Malaria* (*M*)*:* variable que se refiere a la presencia o ausencia de dicha enfermedad. Se denotará la presencia de malaria con $+m$ , y su ausencia con $\neg m$ .
- *Grupo sanguíneo* (*S*): variable que se refiere al nivel de inmunidad asociada los dos grupos sanguíneos. Se denota con $s_+$ el grupo que presenta mayor inmunidad a la enfermedad, y con $s_-$ el grupo que presenta menor inmunidad.
- *Región* (*R*): variable que se refiere al nivel de riesgo de contagio asociado a cada una de las tres regiones. Se denota con $r_a$ a la región que presenta un alto riesgo, con $r_m$ a la región con un riesgo medio y con $r_b$ a la región que presenta un bajo riesgo de contagio.

La siguiente tabla presenta los valores asociados a la probabilidad conjunta $P(m, r, s)$ :

| $P(m, r, s)$ | $+m$ | | | $\neg m$ | | |
|---|---|---|---|---|---|---|
| | $r_a$ | $r_m$ | $r_b$ | $r_a$ | $r_m$ | $r_b$ |
| $s_+$ | 0,0036 | 0,0024 | 0,0042 | 0,0564 | 0,1176 | 0,4158 |
| $s_-$ | 0,0048 | 0,0032 | 0,0056 | 0,0352 | 0,0768 | 0,2744 |

En esta tabla están recogidas las probabilidades marginales $P(r, s)$ .

| $P(r, s)$ | $r_a$ | $r_m$ | $r_b$ | $P(r)$ |
|---|---|---|---|---|
| $s_+$ | 0,060 | 0,120 | 0,420 | 0,600 |
| $s_-$ | 0,040 | 0,080 | 0,280 | 0,400 |
| $P(r)$ | 0,100 | 0,200 | 0,700 | 1,000 |

De esta tabla se obtiene que:

$$P(r_a)=0{,}10 \qquad P(r_m)=0{,}20 \qquad P(r_b)=0{,}70$$

Estas probabilidades indican que los habitantes de la región de alto riesgo representan el 10% de la población, los de la región de riesgo medio el 20%, y los de la región de riesgo bajo el 70%.

| $P(m\|r, s)$ | $+m$ | | | $\neg m$ | | |
|---|---|---|---|---|---|---|
| | $r_a$ | $r_m$ | $r_b$ | $r_a$ | $r_m$ | $r_b$ |
| $s_+$ | 0,06 | 0,02 | 0,01 | 0,94 | 0,98 | 0,99 |
| $s_-$ | 0,12 | 0,04 | 0,02 | 0,88 | 0,96 | 0,98 |

| $P(r, s\|+m)$ | $r_a$ | $r_m$ | $r_b$ | $P(s\|+m)$ |
|---|---|---|---|---|
| $s_+$ | 0,151 | 0,101 | 0,176 | 0,429 |
| $s_-$ | 0,202 | 0,134 | 0,235 | 0,571 |
| $P(r\|+m)$ | 0,353 | 0,235 | 0,412 | 1,000 |

| $P(r, s\|\neg m)$ | $r_a$ | $r_m$ | $r_b$ | $P(s\|\neg m)$ |
|---|---|---|---|---|
| $s_+$ | 0,0578 | 0,1205 | 0,4259 | 0,6040 |
| $s_-$ | 0,0361 | 0,0787 | 0,2811 | 0,3958 |
| $P(r\|\neg m)$ | 0,0938 | 0,1991 | 0,7070 | 1,0000 |

Mientras que las siguientes probabilidades indican la prevalencia de la malaria en las distintas regiones.

$$P(+m|r_a)=0{,}084 \qquad P(+m|r_m)=0{,}028 \qquad P(+m|r_b)=0{,}014$$

## 2.1.2 Definición

El teorema de Bayes es el resultado fundamental sobre el que vamos a construir todo el conocimiento necesario para la construcción de redes bayesianas. De forma que su asimilación será determinante para la adquisición de los contenidos que se tratarán con posterioridad. Para comprender el teorema de Bayes es necesario conocer dos conceptos fundamentales dentro de la teoría de la Probabilidad, la probabilidad condicionada y el teorema de la probabilidad total.

- **Probabilidad condicionada**

  La probabilidad de que acontezca un suceso *A* cuando se ha producido un suceso *B* se llama probabilidad de *A* condicionada a *B*, y se denota $P(A|B)$ .

$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

  Esta definición tiene sentido siempre que $P(B)>0$ .

- **Teorema de la probabilidad total**

  Si un conjunto formado por los sucesos $A_i\subset\Omega,\ i=1,2,...n,$ verifica:

  - Su unión es el suceso seguro (*S*), $\overset{n}{\underset{i=1}{U}} A_i=S$
  - Los sucesos son incompatibles entre sí, $A_i\cap A_j \ \ con \ \ i\neq j$

  Este conjunto se denomina conjunto completo de sucesos y constituye una partición del espacio muestral $\Omega$ .

  Si además, este conjunto de sucesos cumplen que $P(A_i)>0 \ \ \forall i$ , entonces se puede enunciar el teorema de la probabilidad total que dice:

  Para cualquier suceso $B\subset\Omega$ , podemos calcular $P(B)$ haciendo uso de la siguiente expresión:

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

Una vez que se han repasado la probabilidad condicionada y el teorema de la probabilidad total, la definición del teorema de Bayes no presentará dificultad alguna.

Dado un conjunto completo de sucesos $A_i \subset \Omega$, $i = 1, 2, ... n$, un suceso cualquiera $B \subset \Omega$, y considerando que se conocen todas las probabilidades $P(B|A_i)$, el **teorema de Bayes** establece que:

$$P(A_j|B) = \frac{P(B|A_j) \cdot P(A_j)}{P(B)} \quad para\ cualquier\ j = 1, ..., n$$

Si se aplica al denominador el teorema de la probabilidad total se obtiene:

$$P(A_j|B) = \frac{P(B|A_j) \cdot P(A_j)}{P(B)} = \frac{P(B|A_j) \cdot P(A_j)}{\sum_{i=1}^{n} P(B|A_i) \cdot P(A_i)}$$

Esta definición se puede "traducir" con enorme facilidad al contexto de las variables aleatorias, considerando los distintos valores que pueden tomar las variables aleatorias tratadas como los sucesos de la fórmula. De manera que la anterior expresión quedaría de la siguiente forma:

$$P(X = x|Y = y) = \frac{P(Y = y|X = x) \cdot P(X = x)}{P(Y = y)} = \frac{P(Y = y|X = x) \cdot P(X = x)}{\sum_{x'} P(Y = y|X = x') \cdot P(X = x')}$$

donde $x'$ recorre todos los valores que puede tomar la variable $X$.

Para facilitar el manejo y la comprensión, la anterior expresión se puede simplificar mostrando únicamente los valores que toman las variables $X$ e $Y$, ya que se sobreentienden las variables de las que proceden.

$$P(x|y) = \frac{P(y|x) \cdot P(x)}{P(y)} = \frac{P(y|x) \cdot P(x)}{\sum_{x'} P(y|x') \cdot P(x')}$$

**Ejemplo**: para calcular la probabilidad de que un paciente provenga de una determinada región sabiendo que padece malaria se hace uso del teorema de Bayes. La expresión utilizada en dicho cálculo será:

$$P(r|+m) = \frac{P(+m|r) \cdot P(r)}{P(+m|r_a) \cdot P(r_a) + P(+m|r_m) \cdot P(r_m) + P(+m|r_b) \cdot P(r_b)}$$

Haciendo uso de las probabilidades calculadas en el supuesto práctico:

$$P(r_a|+m)=\frac{0{,}084\cdot 0{,}10}{0{,}084\cdot 0{,}10+0{,}028\cdot 0{,}20+0{,}014\cdot 0{,}70}=0{,}353$$

$$P(r_m|+m)=\frac{0{,}028\cdot 0{,}20}{0{,}084\cdot 0{,}10+0{,}028\cdot 0{,}20+0{,}014\cdot 0{,}70}=0{,}235$$

$$P(r_b|+m)=\frac{0{,}014\cdot 0{,}70}{0{,}084\cdot 0{,}10+0{,}028\cdot 0{,}20+0{,}014\cdot 0{,}70}=0{,}412$$

La suma de estas tres probabilidades es igual a 1, como era de esperar, ya que cualquiera de los pacientes incluidos en el estudio pertenece a una de las tres regiones consideradas.

### 2.1.3 Tests diagnósticos

Una de las grandes contribuciones del teorema de Bayes a la Medicina, y en concreto a la Epidemiología, son los **tests diagnósticos**.

A continuación, se definirán los indicadores probabilísticos que constituyen la base fundamental de los tests diagnósticos.

***Notación***: *E* denotará a la variable *Enfermedad* que toma dos valores: padece la enfermedad, $+e$ , y no padece la enfermedad, $\neg e$ . Mientras que *H* denotará la variable *Hallazgo* (este hallazgo puede ser un síntoma, el resultado de una prueba diagnóstica o un signo) que toma dos valores: hallazgo presente, $+h$ , y hallazgo ausente, $\neg h$ .

- **Prevalencia**: probabilidad de que un paciente padezca una determinada enfermedad *E*. La prevalencia se expresa como $P(+e)$ .

- **Sensibilidad**: probabilidad de que se presente un hallazgo *H* cuando un paciente padece una determinada enfermedad *E*. La sensibilidad se expresa como $P(+h|+e)$ . Este indicador informa acerca de la probabilidad de obtener un verdadero positivo, asociado al hallazgo *H*, cuando el paciente sufre la enfermedad *E*.

  Es importante tener en cuenta que una elevada sensibilidad "significa" que la enfermedad casi siempre genera o produce el hallazgo en cuestión. Además, obtenemos una información valiosa al calcular el complementario de este indicador, $1-P(+h|+e)=P(\neg h|+e)$ , que se corresponde con la probabilidad de obtener un falso negativo cuando se padece la enfermedad. De modo que, un test o prueba que sea muy sensible ofrecerá pocos falsos negativos. Esto implica que un negativo

obtenido con una prueba o test muy sensible será bastante fiable. Así, las pruebas o tests con alta sensibilidad son de gran utilidad para descartar la presencia de una determinada enfermedad.

- **Especificidad**: probabilidad de que no se presente un hallazgo *H* cuando el paciente no padece la enfermedad. Se expresa como $P(\neg h|\neg e)$ . Un test que presente una alta especificidad "significa" que resulta bastante improbable que la presencia del hallazgo *H* (asociado al test) sea producida por causas distintas de la enfermedad en cuestión. El complementario de la especificidad, $P(+h|\neg e)=1-P(\neg h|\neg e)$ , sería muy bajo en el caso de una especificidad elevada, y éste informa acerca de la probabilidad de obtener un falso positivo cuando la enfermedad no está presente en el paciente. Con lo cual, un positivo asociado al hallazgo aporta una fuerte evidencia en favor de la presencia de la enfermedad. De ahí que, cuando se desea "confirmar" la presencia de una enfermedad se haga uso de pruebas con un alto grado de especificidad.

- **Valor predictivo positivo (VPP)**: probabilidad de que la enfermedad *E* sea padecida por un paciente que presenta el hallazgo *H*. Este indicador se expresa como $P(+e|+h)$ . El valor predictivo positivo se calcula haciendo uso del teorema de Bayes y de los indicadores probabilísticos presentados anteriormente.

$$P(+e|+h)=\frac{P(+h|+e)\cdot P(+e)}{P(+h|+e)\cdot P(+e)+P(+h|\neg e)\cdot P(\neg e)}=\frac{P(+h|+e)\cdot P(+e)}{P(+h|+e)\cdot P(+e)+(1-P(\neg h|\neg e))\cdot(1-P(+e))}$$

La última expresión se puede escribir en función de la prevalencia (*Preval*), la sensibilidad (*Sensib*) y la especificidad (*Especif*):

$$P(+e|+h)=VPP=\frac{Sensib\cdot Preval}{Sensib\cdot Preval+(1-Especif)\cdot(1-Preval)}$$

- **Valor predictivo negativo (VPN)**: probabilidad de que la enfermedad *E* no esté presente en un paciente que no presenta el hallazgo *H*. Este indicador se expresa como $P(\neg e|\neg h)$ . El valor predictivo negativo se calcula haciendo uso del teorema de Bayes y de los indicadores probabilísticos presentados anteriormente.

$$P(\neg e|\neg h)=\frac{P(\neg h|\neg e)\cdot P(\neg e)}{P(\neg h|+e)\cdot P(+e)+P(\neg h|\neg e)\cdot P(\neg e)}=\frac{P(\neg h|\neg e)\cdot P(\neg e)}{(1-P(+h|+e))\cdot P(+e)+P(\neg h|\neg e)\cdot(1-P(+e))}$$

La última expresión se puede escribir en función de la prevalencia (*Preval*), la sensibilidad (*Sensib*) y la especificidad (*Especif*):

$$P(\neg e|\neg h) = VPN = \frac{Especif \cdot (1 - Preval)}{(1 - Sensib) \cdot Preval + Especif \cdot (1 - Preval)}$$

**Ejemplo**: Si una enfermedad *E* presenta una prevalencia de 0,001, y se dispone de un test *H* con una sensibilidad del 98% (0,98) y una especificidad del 96% (0,96). Los valores predictivos asociados a estos valores serían:

$$P(+e|+h) = VPP = \frac{0{,}98 \cdot 0{,}001}{0{,}98 \cdot 0{,}001 + 0{,}04 \cdot 0{,}999} = 0{,}024$$

$$P(\neg e|\neg h) = VPN = \frac{0{,}96 \cdot 0{,}999}{0{,}02 \cdot 0{,}001 + 0{,}96 \cdot 0{,}999} = 0{,}999979$$

El valor obtenido para el *VPP* indica que aunque el test *H* ofrezca un resultado positivo aún existe una pequeña probabilidad de que el paciente en cuestión padezca la enfermedad *E*. La razón de ésto radica en que para una prevalencia tan baja (0,001) se necesita una especificidad más elevada para "confirmar" la presencia de la enfermedad.

Mientras que el elevado valor obtenido para el *VPN* indica que si el test arroja un resultado negativo estamos casi seguros (99,9979% de certeza) de que el paciente en cuestión no padece la enfermedad. Este resultado se debe a la baja prevalencia y a la elevada sensibilidad del test (0,98).

Al hilo de este ejemplo, se puede hablar de **probabilidad a priori**, para definir aquella probabilidad de la que se disponía antes de realizar el test, $P(+e)$ , que se correspondería con la prevalencia (0,001). Y se puede denominar **probabilidad a posteriori** a aquella obtenida tras la realización del test, $P(+e|h)$ .

Para finalizar este apartado dedicado a los tests diagnósticos conviene señalar que mientras la sensibilidad y la especificidad son dos indicadores probabilísticos independientes, los valores predictivos dependen tanto de la sensibilidad como de la especificidad.

## 2.1.4 Expresiones del Teorema de Bayes

En un apartado previo se presentó el teorema de Bayes para dos variables *X* e *Y* en su forma clásica:

$$P(x|y)=\frac{P(y|x)\cdot P(x)}{P(y)}=\frac{P(y|x)\cdot P(x)}{\sum_{x'} P(y|x')\cdot P(x')}$$

Pero esta expresión no es única. Así, el teorema de Bayes se puede expresar, dependiendo del uso que se quiera hacer del mismo, de dos formas alternativas: forma normalizada y forma racional.

**Forma normalizada del teorema de Bayes**

Si se observa con detenimiento la forma clásica del teorema de Bayes, se comprueba que, para cualquier valor $x$ que tome la variable $X$ , el denominador es siempre el mismo. Este hecho propicia la definición de la constante $\alpha$ como:

$$\alpha=\frac{1}{\sum_{x'} P(y|x')\cdot P(x')}$$

Haciendo uso de $\alpha$ , se reescribe la forma clásica como:

$$P(x|y)=\alpha\cdot P(x)\cdot P(y|x)$$

Esta expresión se denomina forma normalizada del teorema de Bayes.

**Ejemplo**: haciendo uso de la forma normalizada del teorema de Bayes se pueden calcular las probabilidades $P(r_a|+m)$ , $P(r_m|+m)$ y $P(r_b|+m)$ .

$$P(r_a|+m)=\alpha\cdot P(r_a)\cdot P(+m|r_a)=\alpha\cdot 0{,}10\cdot 0{,}084=0{,}0084\cdot\alpha$$

$$P(r_m|+m)=\alpha\cdot P(r_m)\cdot P(+m|r_m)=\alpha\cdot 0{,}20\cdot 0{,}028=0{,}0056\cdot\alpha$$

$$P(r_b|+m)=\alpha\cdot P(r_b)\cdot P(+m|r_b)=\alpha\cdot 0{,}70\cdot 0{,}014=0{,}0098\cdot\alpha$$

Para averiguar el valor de las distintas probabilidades es necesario conocer el valor de $\alpha$ . Teniendo en cuenta que la suma de las tres probabilidades ha de ser 1, se plantea la siguiente ecuación:

$$P(r_a|+m)+P(r_m|+m)+P(r_b|+m)=1$$

Sustituyendo las expresiones obtenidas para cada probabilidad:

$$(0{,}0084+0{,}0056+0{,}0098)\cdot\alpha=1$$

$$0{,}0238\cdot\alpha=1\rightarrow\alpha=\frac{1}{0{,}0238}=42{,}017$$

Sustituyendo ahora en cada una de las expresiones obtenidas en primera instancia:

$$P(r_a|+m)=0{,}0084\cdot\alpha=0{,}0084\cdot 42{,}016=0{,}353$$

$$P(r_m|+m)=0{,}0056\cdot\alpha=0{,}0056\cdot 42{,}016=0{,}235$$

$$P(r_b|+m)=0{,}0098\cdot\alpha=0{,}0098\cdot 42{,}016=0{,}412$$

Como se puede observar, haciendo uso de la forma normalizada se han obtenido los mismos resultados que se obtuvieron mediante la forma clásica.

La forma normalizada facilita la interpretación y comprensión del teorema de Bayes. En su expresión se distingue claramente el término asociado a la probabilidad a posteriori $(P(x|y))$ , el término asociado a la probabilidad a priori $(P(x))$ y el asociado a la verosimilitud del hallazgo $y$ cuando se verifica la hipótesis $x$ , $(P(y|x))$ . Este último término permitirá seleccionar aquel diagnóstico $x$ que hace más verosímil (probable) el hallazgo $y$ . Coloquialmente, se podría decir que la verosimilitud “mide” la concordancia de los diagnósticos con determinados hallazgos.

**Forma racional del teorema de Bayes**

Esta expresión del teorema de Bayes es especialmente útil cuando se desean comparar dos diagnósticos $x_1$ y $x_2$ frente a un hallazgo $y$ , de modo que el interés radica en el diagnóstico diferencial, es decir, averiguar cuánto más probable es el diagnóstico $x_1$ frente al diagnóstico $x_2$ , o viceversa.

Haciendo uso de la forma normalizada del teorema de Bayes se tiene:

$$P(x_1|y)=\alpha\cdot P(x_1)\cdot P(y|x_1)$$

$$P(x_2|y)=\alpha\cdot P(x_2)\cdot P(y|x_2)$$

Si se dividen las expresiones anteriores se obtiene:

$$\frac{P(x_1|y)}{P(x_2|y)}=\frac{P(x_1)}{P(x_2)}\cdot\frac{P(y|x_1)}{P(y|x_2)}$$

- La fracción $\frac{P(x_1|y)}{P(x_2|y)}$ se denomina razón de probabilidad a posteriori de $x_1$ y $x_2$ . En inglés, *post-test odds ratio* o *posterior odds*.

- La fracción $\frac{P(x_1)}{P(x_2)}$ se denomina razón de probabilidad a priori. En inglés, *pre-test odds* ratio o *prior odds.*

- La fracción $\frac{P(y|x_1)}{P(y|x_2)}$ se denomina razón de verosimilitud. En inglés, *likelihood ratio*.

**Ejemplo**: si se considera un paciente que padece malaria, ¿cuánto más probable es que su procedencia sea la región de alto riesgo $(r_a)$ frente a que la procedencia alternativa sea la región de riesgo medio $(r_m)$ ?

$$\frac{P(r_a|+m)}{P(r_m|+m)}=\frac{P(r_a)}{P(r_m)}\cdot\frac{P(+m|r_a)}{P(+m|r_m)}=\frac{0{,}10}{0{,}20}\cdot\frac{0{,}084}{0{,}028}=\frac{1}{2}\cdot 3=1{,}5$$

Este resultado indica que: para un paciente que padece malaria es 1,5 veces más probable que pertenezca a la región $r_a$ que a la región $r_m$ .

Como se ha podido observar, la forma racional del teorema de Bayes no nos informa de cuál es la probabilidad de cada diagnóstico, sino que nos proporciona una razón de probabilidad. No obstante, existe un caso concreto en el que se puede calcular la probabilidad de cada diagnóstico en base a la razón de probabilidad: cuando solo existen dos diagnósticos posibles. La situación más común en la que se verifica esta condición es cuando se desea determinar la presencia o ausencia de una enfermedad.

Así, cuando se quiere diagnosticar la presencia $(+e)$ o ausencia $(\neg e)$ de una enfermedad *E*, y se dispone del hallazgo *H*; la forma racional presenta el siguiente aspecto:

$$\frac{P(+e|h)}{P(\neg e|h)}=\frac{P(+e)}{P(\neg e)}\cdot\frac{P(h|+e)}{P(h|\neg e)}$$

La anterior expresión se puede escribir como:

$$RP_{post}(E)=RP_{pre}(E)\cdot RV_E(h)$$

Cada uno de los términos que componen la anterior expresión se pueden calcular del siguiente modo:

- Razón de probabilidad a priori para la enfermedad *E*:

$$RP_{pre}(E)=\frac{P(+e)}{P(\neg e)}=\frac{Preval}{1-Preval}$$

- Razón de probabilidad a posteriori para la enfermedad *E*:

$$RP_{post}(E)=\frac{P(+e|h)}{P(\neg e|h)}=\frac{P(+e|h)}{1-P(+e|h)}$$

- Razón de verosimilitud para *E* frente al valor *h* del hallazgo *H*:

$$RV_{E}=\frac{P(h|+e)}{P(h|\neg e)}$$

Con las razones anteriores se pueden calcular las probabilidades a posteriori haciendo uso de las siguientes expresiones:

$$P(+e|h)=\frac{RP_{post}(E)}{1+RP_{post}(E)} \qquad P(\neg e|h)=\frac{1}{1+RP_{post}(E)}$$

# 2.2 Asociaciones estadísticas y causalidad

El conocimiento de los "mecanismos" subyacentes en los distintos tipos de asociaciones estadísticas entre variables resulta determinante tanto en la construcción de redes bayesianas como en la creación de cualquier sistema de decisión probabilístico.

## 2.2.1 Independencia y correlación

Se dice que hay **correlación positiva** entre dos valores concretos $x$ e $y$ de las variables aleatorias $X$ e $Y$, respectivamente, cuando:

$$P(x,y)>P(x)\cdot P(y)$$

Siempre que $P(y)\neq 0$, la anterior desigualdad se puede transformar en:

$$P(x|y)>P(x)$$

Esta desigualdad revela que, en una situación de correlación positiva entre los valores *x* e *y*, el hecho de conocer que la variable *Y* toma el valor *y* provoca un aumento de la probabilidad de que la variable *X* tome el valor *x*.

Mientras que si $P(x)\neq 0$ , al dividirla por este factor se transforma:

$$P(y|x) > P(y)$$

La interpretación de esta desigualdad es análoga a la anterior.

Es importante señalar que modificando la desigualdad > por < ó = obtenemos las definiciones de **correlación negativa** y **correlación nula**, respectivamente. Donde la correlación negativa implica que el hecho de conocer que la variable *Y* toma el valor *y* provoca un descenso de la probabilidad de que la variable *X* tome el valor *x*, y viceversa. Y la correlación nula implica que el conocimiento de un valor de una variable no afecta a la probabilidad asociada al valor de la otra variable.

La anterior definición de correlación entre dos valores se puede extender, fácilmente, a las dos variables a las que pertenecen los valores.

Se considera que las variables *X* e *Y* están correlacionadas cuando existe correlación, positiva o negativa, entre algún par de valores $(x, y)$ . Cuando no existe ningún par de valores $(x, y)$ que presenten correlación, entonces se dice que las variables *X* e *Y* son variables independientes, en sentido probabilístico; es decir, que conocer el valor que toma una de las variables no modifica la probabilidad asociada a los valores de la otra, $P(x, y)=P(x)\cdot P(y) \ \forall x, y$ .

Habitualmente, podemos encontrar correlación positiva entre una enfermedad y una prueba o test asociado a la misma.

Dos conceptos van a ser de extraordinaria importancia para los fundamentos probabilísticos asociados a las redes bayesianas: independencia y dependencia condicional. Para su definición se utilizan tres variables *X*, *Y* y *Z* , y las ideas expuestas anteriormente (correlación positiva, correlación negativa, etc.) en su versión "condicional".

Se dice que hay **correlación positiva dado el valor *z*** (de la variable $Z$ ) entre dos valores concretos $x$ e $y$ de las variables aleatorias $X$ e $Y$ , respectivamente, cuando:

$$P(x, y|z) > P(x|z)\cdot P(y|z)$$

Siempre que $P(y,z)\neq 0$ , la anterior desigualdad se puede transformar en:

$$P(x|y,z)>P(x|z)$$

Esta desigualdad revela que, en una situación de correlación positiva entre los valores *x* e *y* sabiendo que la variable *Z* toma el valor *z*, el hecho de conocer que la variable *Y* toma el valor *y* provoca un aumento de la probabilidad de que la variable *X* tome el valor *x*.

Mientras que si, $P(x,z)\neq 0$ al dividirla por este factor se transforma:

$$P(y|x,z)>P(y|z)$$

La interpretación de esta desigualdad es análoga a la anterior.

Es importante señalar que modificando la desigualdad > por < ó = obtenemos las correspondientes definiciones de **correlación negativa** y **correlación nula dado el valor *z***, respectivamente. Donde la correlación negativa implica que el hecho de conocer que la variable *Y* toma el valor *y* provoca un descenso de la probabilidad de que la variable *X* tome el valor *x*, y viceversa. Y la correlación nula implica que el conocimiento de un valor de una variable no afecta a la probabilidad asociada al valor de la otra variable.

La definición de correlación condicional entre tres valores se puede extender, fácilmente, a las tres variables a las que pertenecen los valores.

Se dice que existe correlación entre las variables *X* e *Y* dada la variable *Z* cuando existe correlación positiva o negativa para alguna terna de valores $(x,y,z)$ .

Dos variables *X* e *Y* no están correlacionadas dada la variable *Z* cuando:

$$P(x,y|z)=P(x|z)\cdot P(y|z)\ \ \forall x,y,z$$

es decir, cuando no existe ningún par de valores $(x,y)$ correlacionados para ningún valor $z$ . En este caso se dice que las variables *X* e *Y* son condicionalmente independientes dada la variable *Z.*

### 2.2.2 Independencia probabilística vs. independencia causal

El proceso de construcción de redes bayesianas lleva implícito una serie de destrezas que resultan de capital importancia para su éxito. Una de éstas es la capacidad de distinguir entre independencia probabilística e independencia causal.

- **Independencia probabilística**: está asociada a la verificación de la igualdad

  $$P(x, y) = P(x) \cdot P(y) \quad \forall x, y$$

  En esta situación no existe correlación entre ningún par de valores de *X* e *Y* (correlación nula).

- **Independencia causal**: implica que el valor tomado por una variable no ejerce influencia sobre el valor que toma la otra variable.

Es importante reseñar que, generalmente, la independencia causal implica independencia probabilística, del mismo modo que la dependencia causal suele implicar correlación. A menudo, se suele incurrir en el error de pensar en la existencia de una relación de causalidad entre dos variables correlacionadas. Sin embargo, la correlación no tiene porque implicar causalidad. Y de hecho, una de las tareas más difíciles es la constatación de una relación de causalidad entre dos variables.

### 2.2.3 ¿Causalidad = Correlación?

Una de las cuestiones más controvertidas dentro del ámbito probabilístico es si la correlación nos ofrece una información que nada tiene que ver con la causalidad o si en determinadas condiciones la correlación puede tener connotaciones causales. En realidad, la cuestión de fondo es la definición y naturaleza de la causalidad. De cualquier modo, resulta incuestionable que la causalidad es un concepto distinto de la correlación, y que se podría decir que está en un nivel superior en lo que a asociación entre variables se refiere.

Lo más importante, es tener en cuenta que la causalidad implica correlación, pero que el recíproco no es cierto.

En el siguiente ejemplo se puede comprobar claramente como la correlación no implica la causalidad.

**Ejemplo**: Podría darse el caso de que al realizar un estudio de distintos aspectos relacionados con una ciudad se encuentre que existe correlación entre el número de máquinas tragaperras y el número de nacimientos. Es obvio que entre estas dos variables no existe una relación de causalidad. Con lo cual se verifica la afirmación realizada anteriormente (correlación no implica causalidad). Pero, ¿a qué se debe la correlación entre estas dos variables?

El número de habitantes resulta determinante para el número de bares, de modo que, normalmente, a mayor número de habitantes mayor cantidad de bares. Y los bares son los locales donde se instalan, habitualmente, las tragaperras. Así, queda a la vista la correlación entre el número de habitantes y el de tragaperras. Por otro lado, resulta obvia la correlación entre el número de habitantes y nacimientos. La clave de esta situación está en las variables *Número de habitantes* y *Número de bares* que sirven de "puente" en el establecimiento de la correlación entre el *Número de tragaperras* y el *Número de nacimientos*. Las variables *Número de habitantes* y *Número de bares* son variables de confusión, porque inducen una correlación entre las variables *Número de tragaperras* y *Número de nacimientos* que conduce a equívocos causales. La siguiente figura ilustra esta explicación.

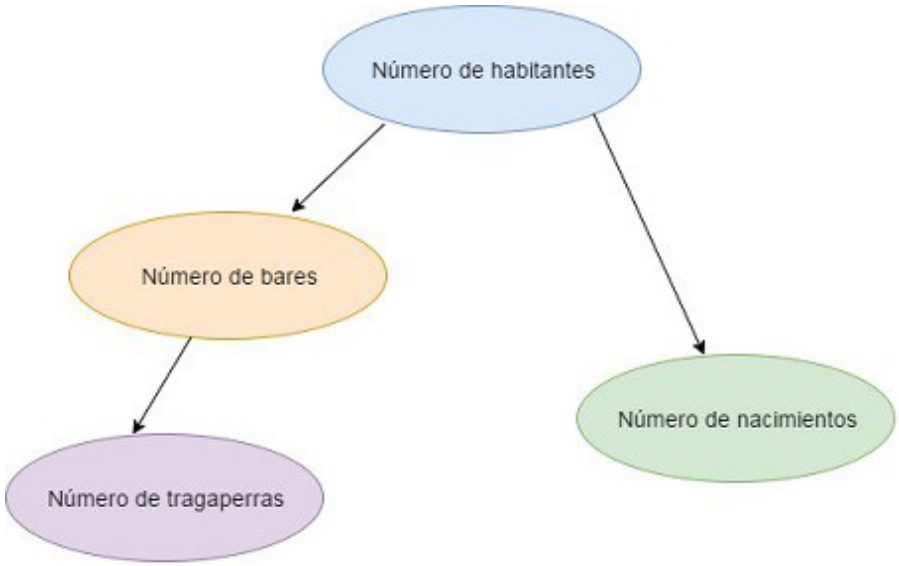



*Figura 2: Correlación entre nº nacimientos y nº de tragaperras*

Cuando en un estudio se detecta una correlación "anómala" entre dos variables, se intentará identificar la variable o variables de confusión que puedan estar interviniendo en este evento. Una vez identificadas e integradas en el estudio, la correlación desaparecerá, de modo que las variables que estaban correlacionadas pasarán a ser condicionalmente independientes.

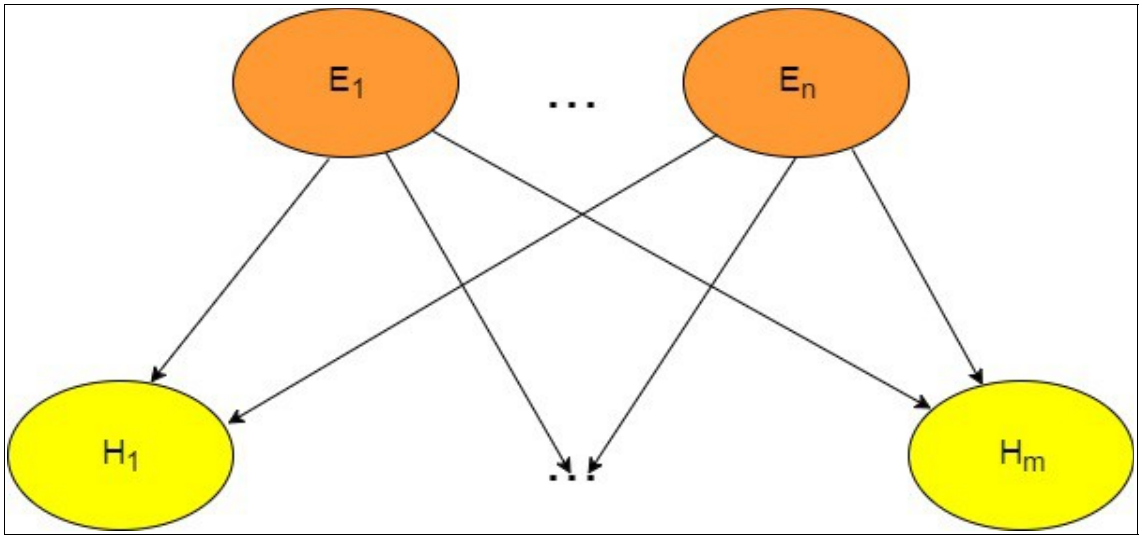
# 2.3 Redes bayesianas

## 2.3.1 Diagnóstico probabilístico

Las redes bayesianas constituyen el refinamiento de un método de diagnóstico probabilístico conocido como el método probabilístico clásico.

Para entender mucho mejor los fundamentos probabilísticos subyacentes en las redes bayesianas, resulta conveniente analizar, en primer lugar, cómo funciona el método probabilístico clásico y cuáles son sus ventajas e inconvenientes.

Se quieren diagnosticar $n$ enfermedades $E_1, E_2, E_3, \ldots, E_n$ a partir de un conjunto de $m$ hallazgos $H_1, H_2, H_3, \ldots, H_n$ . Las implicaciones causales de este planteamiento se pueden observar en el siguiente grafo.




El objetivo de esta situación sería calcular probabilidades del tipo:

$$P(e_1, e_2, \ldots, e_n | h_1, h_2, \ldots, h_m)$$

donde $e_1, e_2, \ldots, e_n$ representan las $n$ enfermedades y $h_1, h_2, \ldots, h_m$ los $m$ hallazgos considerados en el diagnóstico.

Estas probabilidades se refieren a la probabilidad de que cada una de las distintas enfermedades esté presentes o ausentes cuando se conoce una serie de observaciones asociadas a cada uno de los distintos hallazgos.

Para calcular $P(e_1, e_2, \ldots, e_n | h_1, h_2, \ldots, h_m)$ es necesario conocer la siguiente información:

- Probabilidad a priori de las distintas enfermedades, $P(e_1, e_2, \ldots, e_n)$ .
- Probabilidad de que se obtengan determinadas observaciones asociadas a los distintos hallazgos conocida la presencia o ausencia de las distintas enfermedades, $P(h_1, h_2, \ldots, h_m | e_1, e_2, \ldots, e_n)$ .

Integrando todas la probabilidades en el teorema de Bayes:

$$P(e_1, e_2, \ldots, e_n | h_1, h_2, \ldots, h_m) = \frac{P(h_1, h_2, \ldots, h_m | e_1, e_2, \ldots, e_n) \cdot P(e_1, e_2, \ldots, e_n)}{\sum_{e'_1, \ldots, e'_n} P(h_1, h_2, \ldots, h_m | e'_1, e'_2, \ldots, e'_n) \cdot P(e'_1, e'_2, \ldots, e'_n)}$$

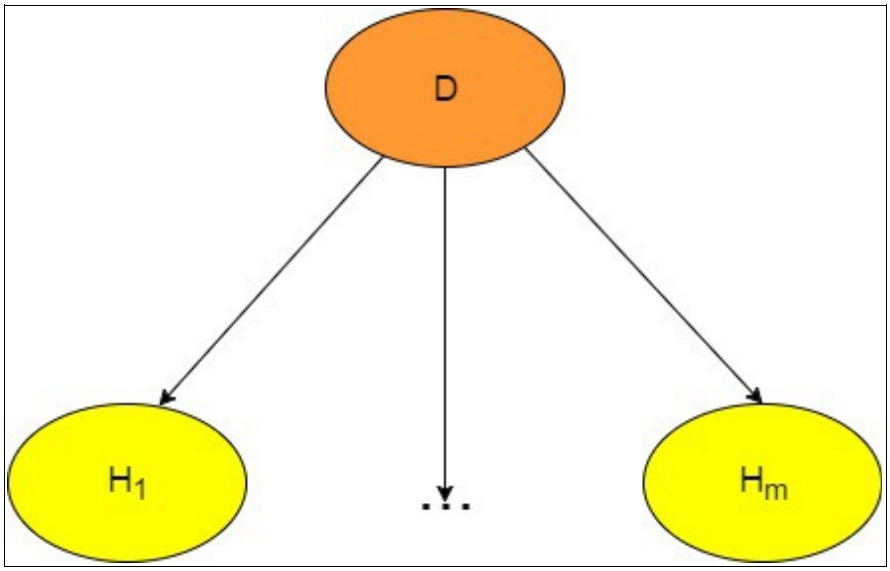
Este método de diagnóstico probabilístico presenta un gran *handicap* en la práctica: la ingente cantidad de información que demanda. Considerando el caso más sencillo, que las variables asociadas tanto a las enfermedades como a los hallazgos sean binarias, se tendrían las siguientes "necesidades":

- ✔ $2^n$ probabilidades a priori del tipo $P(e_1, e_2, \ldots, e_n)$ . Como la suma de estas probabilidades es 1, son suficientes $2^n - 1$ "parámetros" (probabilidades) independientes para expresar todas las probabilidades.
- ✔ $2^{m+n}$ Probabilidades del tipo $P(h_1, h_2, \ldots, h_m | e_1, e_2, \ldots, e_n)$ . El hecho de que estas probabilidades condicionadas sumen 1, hace que sean suficientes $(2^m - 1) \cdot 2^n$ "parámetros" independientes para expresar todas estas probabilidades.

De manera que, en total se necesitarían un total de $2^{m+n} - 1$ "parámetros" independientes para expresar todas las probabilidades implicadas en el modelo. Así, en el caso concreto en el que queramos diagnosticar 4 enfermedades basándonos en 12 hallazgos de naturaleza binaria se necesitarían 65535 "parámetros".

Para salvar esta enorme dificultad se introdujo una hipótesis que operara una reducción importante en el número de información necesaria para la construcción del modelo, y que estuviera fundamentada en la práctica: el conjunto de diagnósticos considerado ha de ser exhaustivo y excluyente. Es decir, un paciente presentará uno y solo uno de los diagnósticos considerados.

Este planteamiento permite sustituir las $n$ variables asociadas a las distintas enfermedades por una única variable *Diagnóstico* (D) que toma $n+1$ valores: $n$ valores correspondientes a las distintas enfermedades que se desean diagnosticar y un valor adicional para representar la ausencia de las enfermedades consideradas. Las implicaciones causales asociadas a este nuevo enfoque se pueden asimilar fácilmente observando el siguiente grafo.




La sustitución de las $n$ variables asociadas a las enfermedades por una única variable *D* que aglutine todos los diagnósticos también genera cambios en la anterior expresión del teorema de Bayes.

$$P(d|h_1,h_2,...,h_m)=\frac{P(h_1,h_2,...,h_m|d)\cdot P(d)}{\sum_{d'} P(h_1,h_2,...,h_m|d')\cdot P(d')}$$

Con este nuevo paradigma la cantidad de información necesaria para su aplicación se ha reducido notablemente. Considerando la situación en la que las variables asociadas a los hallazgos son binarias se tiene que:

- ✔ $n+1$ probabilidades a priori del tipo $P(d)$ , asociadas a la variable *Diagnóstico* (D). Como han de sumar 1, son suficientes $n$ "parámetros" (probabilidades) independientes para expresar todas las probabilidades.
- ✔ $2^m\cdot(n+1)$ Probabilidades del tipo $P(h_1,h_2,...,h_m|d)$ . El hecho de que estas probabilidades condicionadas sumen 1, hace que sean suficientes $2^m\cdot(n+1)-(n+1)$ "parámetros" independientes para expresar todas estas probabilidades.

En total, se necesitarían $2^m \cdot (n+1) - 1$ "parámetros" independientes para expresar todas las probabilidades implicadas en el modelo. Así, en el caso concreto en el que queramos diagnosticar 4 enfermedades basándonos en 12 hallazgos de naturaleza binaria se necesitarían 20479 "parámetros". Esta cantidad es muy inferior a la obtenida en el modelo previo, aunque sigue siendo muy elevada.

Con el objetivo de reducir aún más la cantidad de información necesaria para aplicar el modelo se implementó una hipótesis adicional: la probabilidad asociada a un determinado hallazgo es probabilísticamente independiente de los resultados obtenidos acerca de los hallazgos restantes. En definitiva, la introducción de la hipótesis de independencia condicional para los hallazgos.

Analíticamente, esta hipótesis consiste en:

$$P(h_1, h_2, ..., h_m | d) = P(h_1 | d) \cdot P(h_2 | d) \cdot ... \cdot P(h_m | d) \quad \forall d$$

Al "integrar" esta condición en la expresión del teorema de Bayes se obtiene:

$$P(d | h_1, h_2, ..., h_m) = \frac{P(h_1 | d) \cdot P(h_2 | d) \cdot ... \cdot P(h_m | d) \cdot P(d)}{\sum_{d'} P(h_1 | d') \cdot P(h_2 | d') \cdot ... \cdot P(h_m | d') \cdot P(d')}$$

Con este nuevo enfoque y considerando de nuevo hallazgos de naturaleza binaria se necesitarían:

- ✔ $n+1$ probabilidades a priori del tipo $P(d)$ , asociadas a la variable *Diagnóstico* (D). Como han de sumar 1, son suficientes $n$ "parámetros" (probabilidades) independientes para expresar todas las probabilidades.
- ✔ $(2 \cdot (n+1)) \cdot m$ probabilidades del tipo $P(h_i | d)$ . El hecho de que estas probabilidades condicionadas sumen 1, hace que sean suficientes $(n+1) \cdot m$ "parámetros" independientes para expresar todas estas probabilidades.

En total, se necesitarían $n + (n+1) \cdot m$ "parámetros" independientes para expresar todas las probabilidades implicadas en el modelo. Así, en el caso concreto en el que queramos diagnosticar 4 enfermedades basándonos en 12 hallazgos de naturaleza binaria se necesitarían 64 "parámetros". Esta cantidad es considerablemente inferior a la obtenida en el modelo previo. Por lo tanto, este paradigma probabilístico permite abordar la construcción de modelos de cierto tamaño.

La expresión del teorema de Bayes asociada a este nuevo modelo es:

$$P(d|h_1, h_2, ..., h_m) = \frac{P(h_1|d) \cdot P(h_2|d) \cdot ... \cdot P(h_m|d) \cdot P(d)}{\sum_{d'} P(h_1|d') \cdot P(h_2|d') \cdot ... \cdot P(h_m|d') \cdot P(d')}$$

Dependiendo de la situación, puede resultar de gran utilidad hacer uso de las diferentes expresiones del teorema de Bayes.

**Forma normalizada**

$$P(d|h_1, h_2, ..., h_m) = \alpha \cdot P(d) \cdot P(h_1|d) \cdot P(h_2|d) \cdot ... \cdot P(h_m|d)$$

con

$$\alpha = \frac{1}{\sum_{d'} P(h_1|d') \cdot P(h_2|d') \cdot ... \cdot P(h_m|d') \cdot P(d')}$$

**Forma racional**

$$\frac{P(d_1|h_1, h_2, ..., h_m)}{P(d_2|h_1, h_2, ..., h_m)} = \frac{P(d_1)}{P(d_2)} \cdot \frac{P(h_1|d_1)}{P(h_1|d_2)} \cdot \frac{P(h_2|d_1)}{P(h_2|d_2)} \cdot ... \cdot \frac{P(h_m|d_1)}{P(h_m|d_2)}$$

La anterior expresión no permite calcular la probabilidad de los diferentes diagnósticos, únicamente se puede averiguar la razón de probabilidad a posteriori entre dos diagnósticos $d_1$ $y$ $d_2$ . Aunque, si lo que se desea es diagnosticar la presencia o ausencia de una única enfermedad $E$ , y los valores $+e$ $y$ $\neg e$ ocupan el lugar de $d_1$ $y$ $d_2$ , respectivamente, en la anterior expresión se obtiene:

$$RP_{post}(E) = RP_{pre}(E) \cdot RV_E(h_1) \cdot RV_E(h_2) \cdot ... \cdot RV_E(h_m)$$

donde $RP_{post}(E)$ es la razón de probabilidad a posteriori para la enfermedad $E$ , $RP_{pre}(E)$ es la razón de probabilidad a priori para la enfermedad $E$ , y cada $RV_E(h_i)$ es la razón de verosimilitud para $E$ frente al valor $h_i$ del hallazgo $H_i$ .

La implementación de la forma racional del método clásico mediante la expresión anterior presenta diversas ventajas:

- Facilita la inclusión de nuevos hallazgos al modelo, ya que únicamente hay que incluir la razón de verosimilitud correspondiente $RV_E(h_i)$ .
- Permite calibrar la contribución de los distintos hallazgos al diagnóstico considerado atendiendo a su razón de verosimilitud.
- Posibilita una valoración rápida de la evidencia aportada por cada hallazgo para un diagnóstico concreto atendiendo a su razón de verosimilitud.

Las dos principales limitaciones que presenta el método probabilístico clásico derivan de las dos hipótesis en las que se basa (conjunto de diagnósticos exhaustivo y excluyente, e independencia condicional):

- La imposibilidad de realizar diagnósticos múltiples (un paciente no puede presentar dos o más enfermedades al mismo tiempo).
- La inviabilidad de que los hallazgos incluyan tanto efectos como causas asociadas a las patologías estudiadas.

Para superar estas importantes limitaciones dentro del marco práctico del diagnóstico probabilístico surgió el concepto de red bayesiana, un modelo gráfico probabilístico que constituye la generalización del método clásico. El gran potencial de las redes bayesianas reside en su capacidad para abarcar y visualizar las relaciones de causalidad existentes entre las distintas variables implicadas en la situación a modelizar.

### 2.3.2 Conceptos fundamentales

Para comprender la definición de red bayesiana es necesario conocer una serie de conceptos asociados a la teoría de grafos.

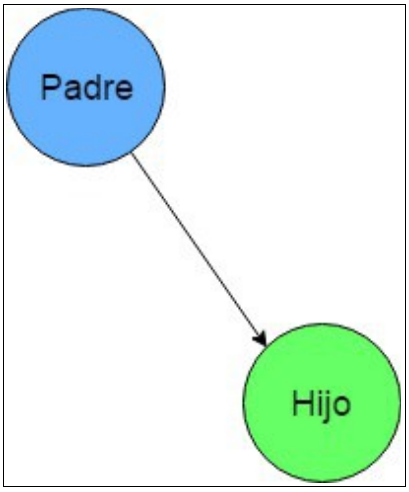
Un grafo es una estructura compuesta por un conjunto de nodos que se encuentran conectados por medio de un conjunto de enlaces.

En lo sucesivo solo se considerarán aquellos grafos cuyos enlaces unen siempre dos nodos diferentes, y en los que no existen dos enlaces que unan un mismo par de nodos. Para definir un enlace se hará uso de las letras que identifican cada nodo. Así, un enlace que une el nodo *A* con el nodo *B*, lo denotaremos con (*A*,*B*). Si el orden de los nodos es relevante entonces el enlace es dirigido. La representación gráfica de un enlace dirigido es una flecha.

Dentro de un grafo dirigido se tienen relaciones de “parentesco”. En un enlace dirigido entre dos nodos, el nodo de “partida” se considera el *padre* del nodo de “llegada” (*hijo*). De manera que a partir de este idea se generan los siguientes conceptos:

- Una familia de nodos está formada por un *hijo* y sus *padres*.

- Un nodo *A* es *antepasado* de un nodo *B* si es su *padre* o si existe un nodo *C* que sea hijo de *A* y *antepasado* de *B.* Si *A* es *antepasado* de *B*, se dice que *B* es *descendiente* de *A*.




Un camino dentro de un grafo es una sucesión de nodos que verifica que cada dos nodos consecutivos están unidos por un enlace. Si los nodos de inicio y fin de un camino coinciden, se dice que el camino es cerrado. Conviene remarcar que en estas dos definiciones en ningún momento se ha considerado la dirección de los enlaces.

Si en un grafo dirigido se puede recorrer un camino cerrado siguiendo la dirección de los enlaces entonces se tiene un ciclo. A un grafo dirigido que no tiene ciclos se le llama grafo dirigido acíclico.

Un grafo es conexo cuando se puede "trazar", al menos, un camino entre dos nodos cualesquiera del mismo.

Gracias a esta breve introducción ya se puede entender en qué consiste un grafo dirigido acíclico conexo, un concepto clave en la definición de red bayesiana.

Una **red bayesiana** es un modelo gráfico probabilístico que presenta las siguientes componentes:

- Un conjunto de variables $X_i$ , que toman una serie de valores. Sus valores se representarán a través de la notación $x_i$ .
- Un grafo dirigido acíclico conexo, de manera que cada nodo que lo compone representa unívocamente a cada una de las variables $X_i$ .

- Una distribución de probabilidad asociada a cada $X_i$ . Para cada una de las variables (nodos) se define una distribución de probabilidad condicionada por la configuración de sus padres en el grafo, $P(x_i|padres(x_i))$ .

  Si el nodo asociado a la variable $X_j$ no tiene padres, entonces:

  $$P(x_j|padres(x_j))=P(x_j)$$

  La probabilidad conjunta de las variables (nodos) que componen la red bayesiana $P(x_1, x_2, \ldots, x_n)$ se calcula:

  $$P(x_1, x_2, \ldots, x_n)=P(x_1|padres(x_1))\cdot P(x_2|padres(x_2))\cdot \ldots \cdot P(x_n|padres(x_n))$$

  De la anterior expresión se infiere la propiedad de Markov, que encierra el potencial de modelización subyacente en las redes bayesianas. Esta propiedad afirma que el nodo asociado a la variable $X_i$ es condicionalmente independiente de sus no descendientes $Z_1, \ldots, Z_m$ dada la configuración de sus padres, $padres(x_i)$ :

  $$P(x_i|padres(x_i), z_1, \ldots, z_m)=P(x_i|padres(x_i))$$

  Una de las conclusiones prácticas que se pueden extraer de la propiedad de Markov es que si dos nodos no presentan antepasados comunes en la red entonces son independientes a priori.

### 2.3.3 Modelos canónicos asociados a las redes bayesianas

Uno de los "componentes" esenciales de una red bayesiana son las distribuciones de probabilidad condicionada (a la configuración de sus padres) de cada uno de sus nodos. Con frecuencia surge un problema relativo al cálculo de las probabilidades asociadas a cada una de estas distribuciones: la existencia de un gran número de padres. Esta característica se traduce a nivel probabilístico en un elevado número de parámetros independientes para definir la distribución. En el caso binario (todos los nodos son binarios), si tuviéramos un nodo *Z* asociado a una enfermedad con 12 padres, que bien podrían ser posibles causas de la enfermedad, la tabla de probabilidad asociada a esta situación tendría $2^{12}=4096$ parámetros independientes. Sin duda, una situación difícilmente abordable.

Para salvar este tipo de dificultades se hace uso de una serie de modelos canónicos probabilísticos que permiten economizar y simplificar los cálculos vinculados a la construcción de las distribuciones de probabilidad asociadas a los nodos.

En este apartado se abordarán los modelos canónicos utilizados para el tratamiento de variables binarias, que constituye el caso más común dentro de la práctica médica. Es importante señalar que estos modelos canónicos provienen del campo de la Inteligencia Artificial.

Los modelos canónicos que se abordarán serán las variantes (determinista, probabilística y residual) de la puerta OR (*OR-gate*).

**Variante determinista**

Si en una red bayesiana, un nodo $Z$ presenta dos causas (padres) representadas por los nodos $X_1$ y $X_2$, de manera que la presencia de una de ellas es suficiente para producir el efecto $Z$, y además no existen otras causas que puedan generar $Z$. Entonces, la presencia del efecto $Z$ estará asociada a la presencia de al menos una de las causas mencionadas. Desde un punto probabilístico, esta situación se traduce en:

| $P(+z\|x_1, x_2)$ | $+x_1$ | $\neg x_1$ |
|---|---|---|
| $+x_2$ | 1 | 1 |
| $\neg x_2$ | 1 | 0 |

Este modelo recibe el apelativo “determinista” porque si se conoce el valor de cualquiera de las causas $X_1$ y $X_2$ se conoce el valor del efecto $Z$.

**Variante probabilística**

Resulta relativamente normal en el campo de las Ciencias de la Salud que una causa no siempre genere un determinado efecto. Para modelizar este tipo de situaciones se hará uso de las probabilidades. Sea $p_i$ la probabilidad de que la causa $X_i$ genere el efecto $Z$, no contemplando la posibilidad de que el efecto sea provocado por otras causas. Si, al igual que antes, el efecto $Z$ presenta únicamente dos padres, $X_1$ y $X_2$, se tiene:

$$p_1 = P(+z|+x_1, \neg x_2)$$
$$p_2 = P(+z|\neg x_1, +x_2)$$

La probabilidad complementaria de $p_i$ indica la probabilidad de que estando presente la causa $X_i$ no se produzca $Z$. Se podría decir que $1-p_i$ es la probabilidad de fallo de la causa $X_i$.

Podría darse el caso de que las causas $X_1$ y $X_2$ estén presentes al mismo tiempo, de modo que $Z$ podría haber sido producido por cualquiera de las dos causas. Con lo que se tendría:

$$P(+z|+x_1,+x_2)=p_1+(1-p_1)\cdot p_2$$

La traducción probabilística de este modelo es:

| $P(+z\|x_1,x_2)$ | $+x_1$ | $\neg x_1$ |
|---|---|---|
| $+x_2$ | $p_1+(1-p_1)\cdot p_2$ | $p_2$ |
| $\neg x_2$ | $p_1$ | 0 |

**Variante residual**

En las anteriores variantes de la puerta OR se suponía que todas las causas asociadas a un efecto o enfermedad eran conocidas. Sin embargo, en la práctica puede ocurrir que no se conozcan todas las causas asociadas a un efecto, o que incluso no resulte adecuado incluir en el modelo algunas de ellas, bien por su carácter extraordinario (enfermedades con una prevalencia muy baja) o porque se desconozca la probabilidad asociada a las mismas. De modo que, aunque las causas representadas (explícitas) en el modelo estén ausentes, todavía existe cierta probabilidad de que el efecto se manifieste. Esta probabilidad recibe el nombre de probabilidad residual.

Para integrar la probabilidad residual en el modelo se construirá una variable que “aglutine” aquellas causas no representadas en el mismo (implícitas).

## 2.3.4 Supuestos prácticos

A continuación, se abordará la modelización de dos supuestos prácticos. Para facilitar una comparativa entre el método probabilístico clásico y las redes bayesianas, el primer supuesto se resolverá mediante los dos modelos. Mientras que el segundo supuesto únicamente se modelizará a través de una red bayesiana.

**Supuesto 1.** El gen *G*, que está presente en el 3% de la población, provoca la enfermedad *A* en el 15% de los casos y la enfermedad *B* en el 60%. Entre las personas que no tienen ese gen, la prevalencia de *A* es el 2% y la de *B* el 1%. Existe un síntoma *S* común a ambas: *A* lo produce en el 70% de los casos y *B* en el 90%. Otras causas, distintas de *A* y *B*, lo producen en el 0,5% de las personas.

Es posible detectar la presencia de *A* mediante un análisis de sangre, con una sensibilidad del 93% y una especificidad del 99%.

Las variables que intervienen en este problema son:

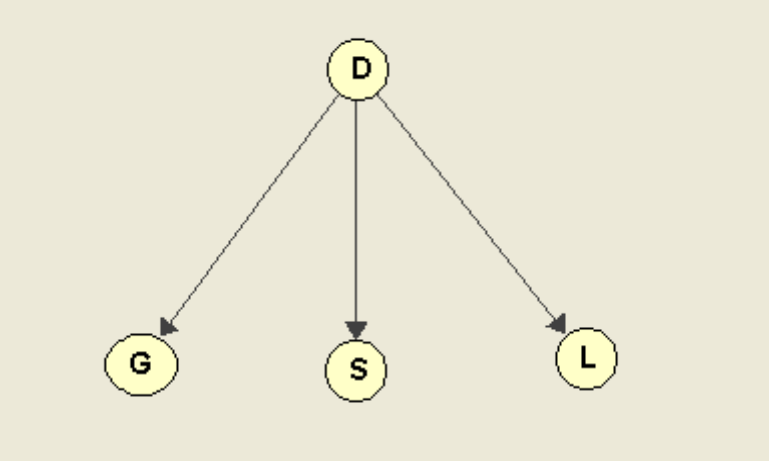
- Variable ***D*** (Diagnóstico): $a\ (enfermedad\ A)$ , $b\ (enfermedad\ B)$ , $n\ (ni\ \ enf.A\ \ ni\, enf.\, B)$

- Variable ***G*** (Test genético): $+g\ (\ presente)$ , $\neg g\ (ausente)$

- Variable ***S*** (Síntoma): $+s\ (\ presente)$ , $\neg s\ (ausente)$

- Variable ***L*** (Análisis de sangre): $+l\ (\ positivo)$ , $\neg l\ (negativo)$

Se construye el grafo asociado al método probabilístico clásico haciendo uso del software *Elvira*.



Se calculan las probabilidades condicionadas que definen el modelo.

Las probabilidades asociadas a la distintos valores que puede tomar la variable *D* se deducen del enunciado del problema.

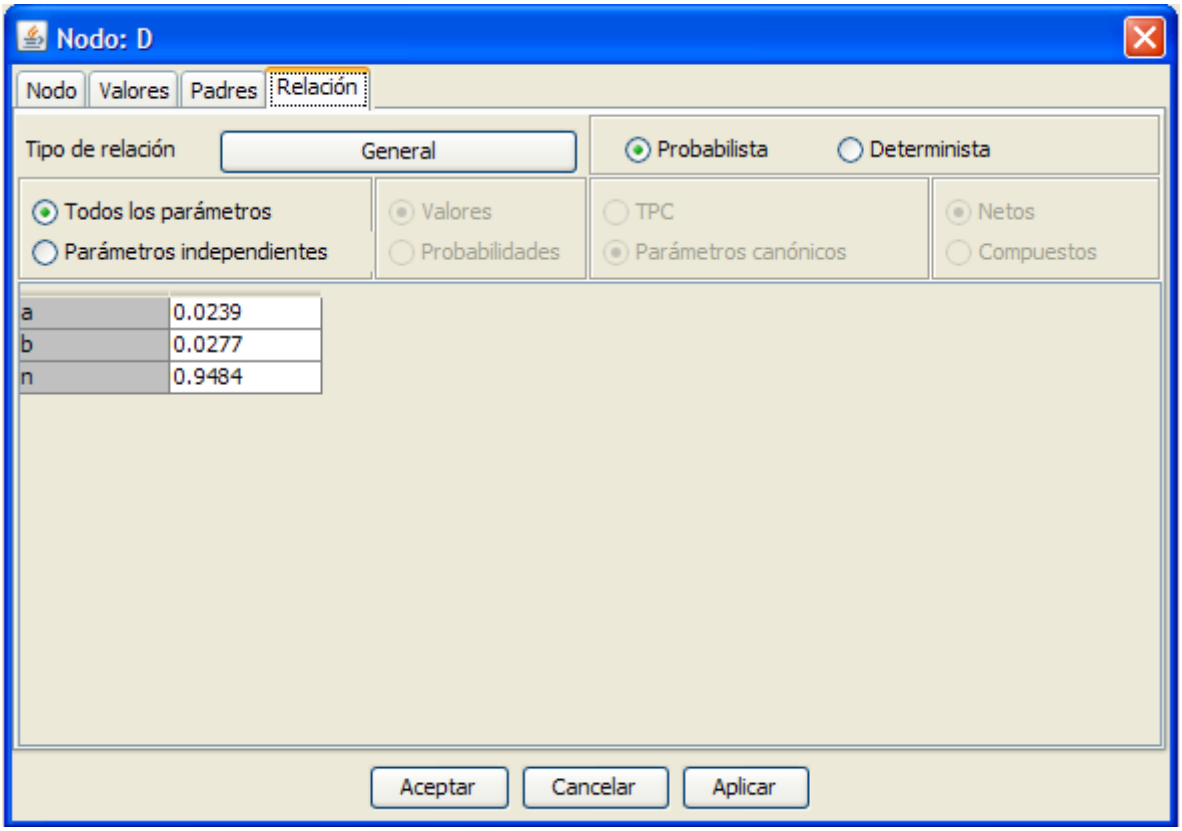
| | $a$ | $b$ | $n$ |
|---|---|---|---|
| $P(D)$ | **0,0239** | **0,0277** | **0,9484** |

$$P(a)=P(a|+g)\cdot P(+g)+P(a|\neg g)\cdot P(\neg g)=0{,}015\cdot 0{,}03+0{,}02\cdot 0{,}97=0{,}0045+0{,}0194=0{,}0239$$

$$P(b)=P(b|+g)\cdot P(+g)+P(b|\neg g)\cdot P(\neg g)=0{,}60\cdot 0{,}03+0{,}01\cdot 0{,}97=0{,}018+0{,}0097=0{,}0277$$

$$P(n)=1-(P(a)+P(b))=1-(0{,}0239+0{,}0277)=0{,}9484$$

Al implementar el problema con *Elvira*, el programa ofrece una tabla de probabilidades similar a la anterior.



La tabla de probabilidades condicionadas:

| $P(G\|D)$ | $a$ | $b$ | $n$ |
|---|---|---|---|
| $+g$ | **0,188** | **0,6498** | **0,00791** |
| $\neg g$ | **0,812** | **0,3502** | **0,99209** |

A continuación, se explicitan los cálculos realizados para los se ha utilizado el teorema de Bayes.

$$P(+g|a)=\frac{P(a|+g)\cdot P(+g)}{P(a)}=\frac{0,15\cdot 0,03}{0,0239}=\frac{0,0045}{0,0239}=0,188$$

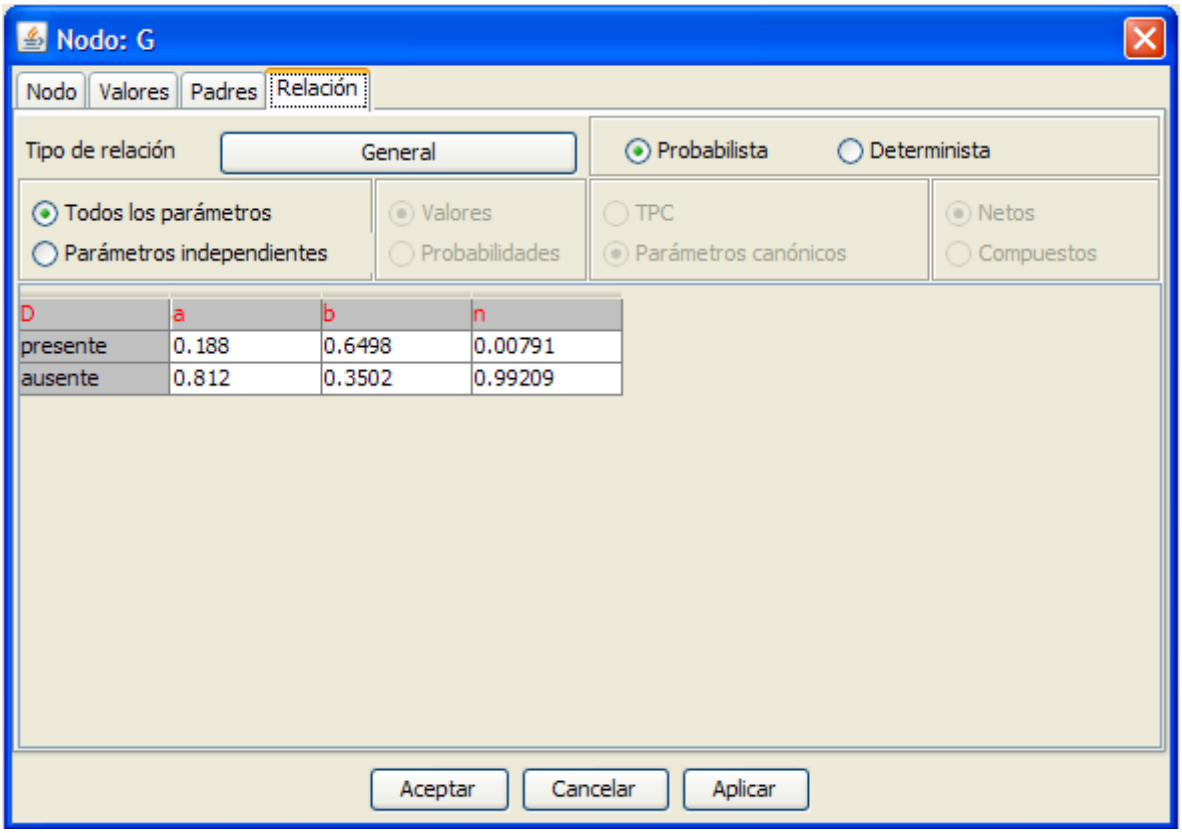
$$P(\neg g|a)=\frac{P(a|\neg g)\cdot P(\neg g)}{P(a)}=\frac{0,02\cdot 0,97}{0,0239}=\frac{0,0194}{0,0239}=0,812$$

$$P(+g|b)=\frac{P(b|+g)\cdot P(+g)}{P(b)}=\frac{0,6\cdot 0,03}{0,0277}=\frac{0,018}{0,0277}=0,6498$$

$$P(\neg g|b)=\frac{P(b|\neg g)\cdot P(\neg g)}{P(b)}=\frac{0,01\cdot 0,97}{0,0277}=\frac{0,0097}{0,0277}=0,3502$$

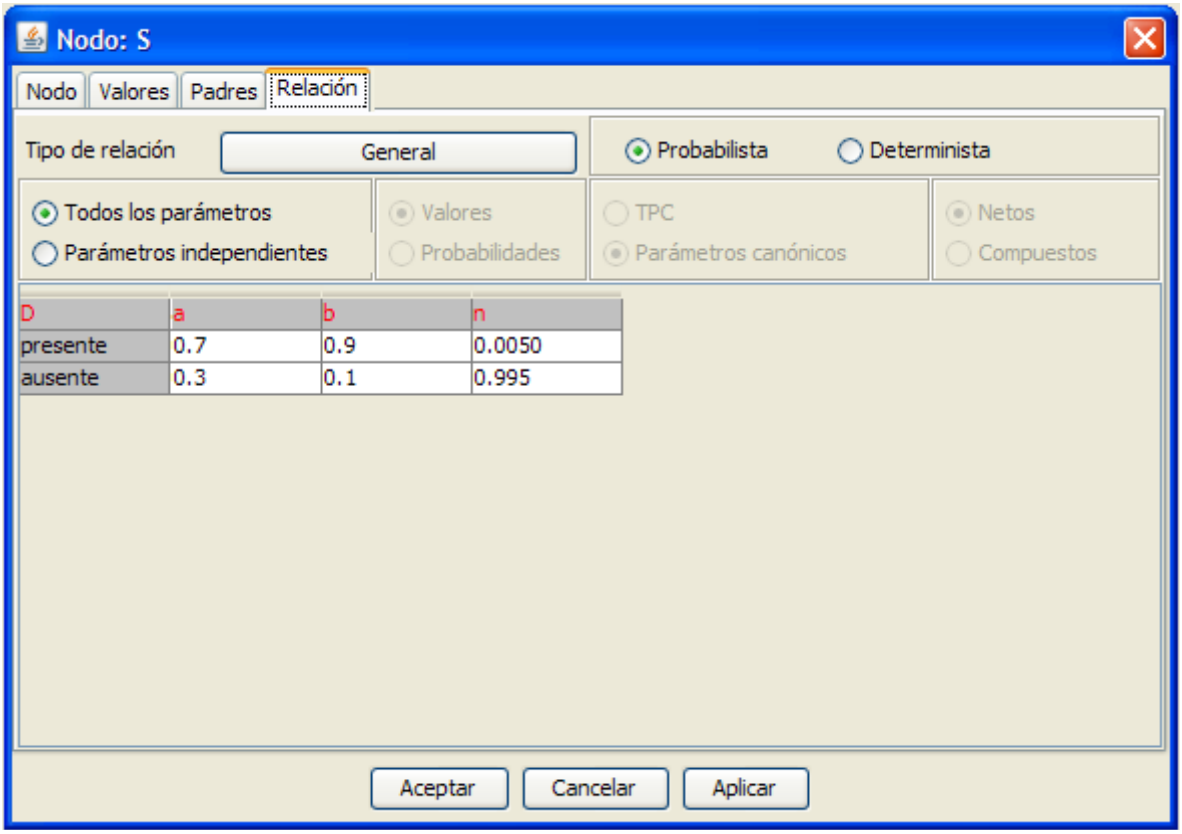
$$P(+g|n)=\frac{P(n|+g)\cdot P(+g)}{P(n)}=\frac{0,25\cdot 0,03}{0,9484}=\frac{0,0075}{0,9484}=0,00791$$

$$P(\neg g|n)=\frac{P(n|\neg g)\cdot P(\neg g)}{P(n)}=\frac{0,97\cdot 0,97}{0,9484}=\frac{0,9409}{0,9484}=0,99209$$




La siguiente tabla se deduce de manera directa del enunciado del problema.

| $P(S\|D)$ | $a$ | $b$ | $n$ |
|---|---|---|---|
| $+s$ | **0,7** | **0,9** | **0,005** |
| $\neg s$ | **0,3** | **0,1** | **0,995** |

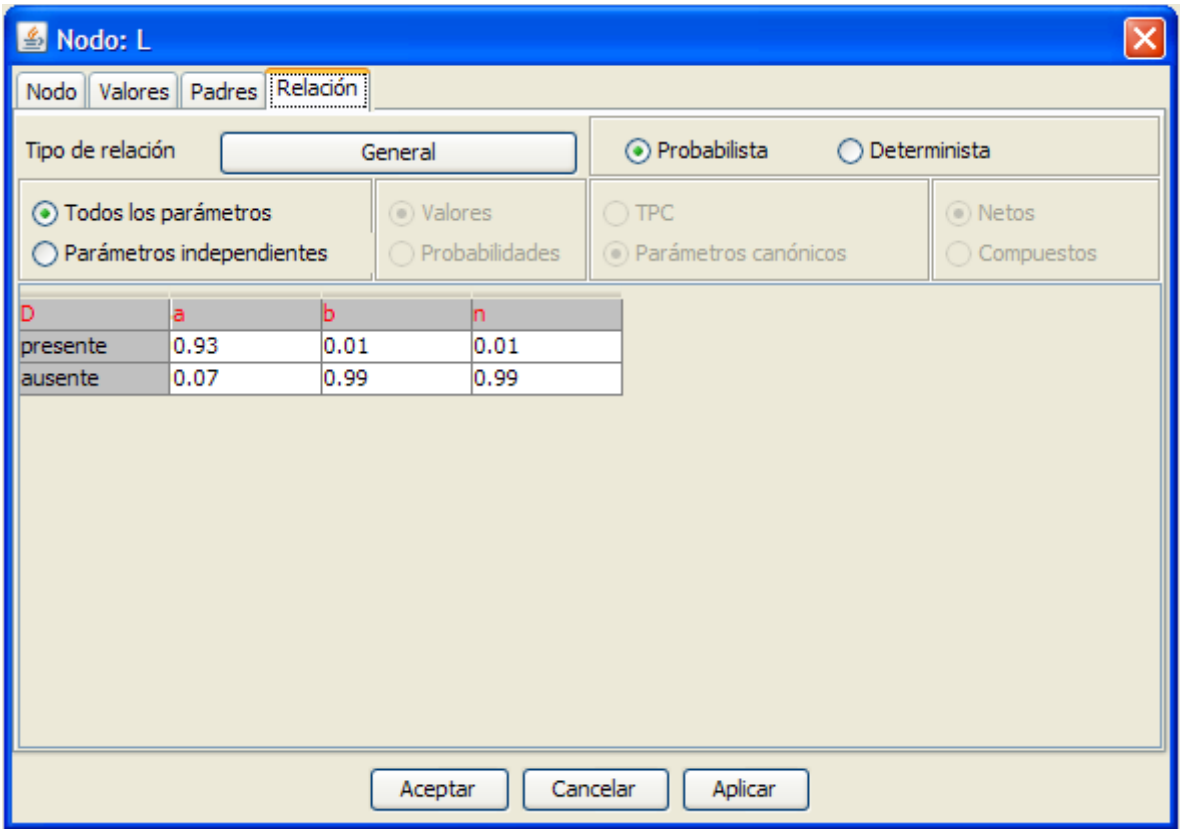



La tabla de probabilidad para *L* es:

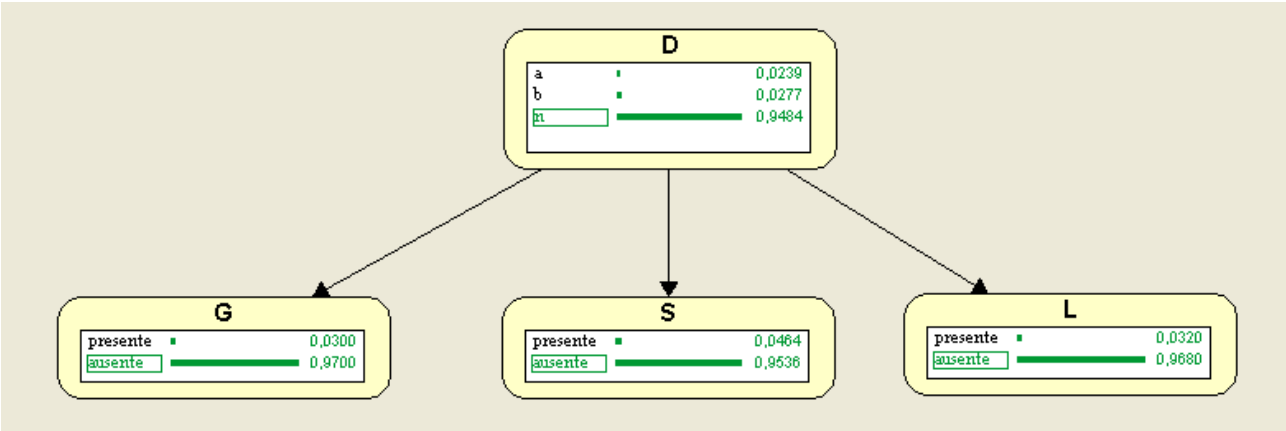
| $P(L\|D)$ | $a$ | $b$ | $n$ |
|---|---|---|---|
| $+l$ | **0,93** | **0,01** | **0,01** |
| $\neg l$ | **0,07** | **0,99** | **0,99** |

En cuanto a la variable *L* (análisis de sangre), del enunciado se desprende que $P(+l\|a)=0{,}93$ y $P(+l\|\neg a)=1-P(\neg l\|\neg a)=1-0{,}99=0{,}01$ . El valor $\neg a$ no es un valor de *D*, y por tanto no sirve directamente para construir la tabla $P(L|D)$ . Lo único que se conoce es que $P(+l\|\neg a)$ es la media ponderada de $P(+l|b)\ y\ P(+l|n)$ . Ahora bien, si se supone que la probabilidad de un falso positivo en la prueba que detecta *A* es la misma para *b* y para *n*, entonces se tiene que estas dos probabilidades son iguales entre si, y por tanto, iguales a su promedio:

$$P(+l|b)=P(+l|n)=P(+l|\neg a)=0{,}01$$




Esta es la imagen que proporciona *Elvira* en el modo *Inferencia* con las probabilidades a priori.



Para construir el modelo se han introducido las dos hipótesis del método probabilista clásico. Es importante analizar si estas hipótesis son razonables.

1ª hipótesis: los diagnósticos son exhaustivos y exclusivos.

Al definir la variable *D* se ha introducido la condición de exhaustividad y exclusividad de los diagnósticos. La condición de exhaustividad se puede comprobar observando los distintos valores que toma la variable *D*, mientras que para comprobar la condición de exclusividad se tienen que hacer una serie de consideraciones. Si las enfermedades *A* y *B* tienen orígenes distintos (independencia causal), serán independientes a priori (en sentido probabilista). Por tanto, la probabilidad de que un paciente padezca a la vez ambas enfermedades es

$$P(A)\cdot P(B)=0{,}0239\cdot 0{,}0277=0{,}00066203\approx\frac{1}{1511}$$

Esto significa que en 1 de cada 1511 pacientes el modelo va a ser incapaz de ofrecer el diagnóstico correcto, pues sólo diagnosticará una de las dos; o dicho de otro modo, que el modelo dará un resultado incorrecto en el 0,066203% del total de pacientes. Si hubiera correlación positiva entre ambas enfermedades, la probabilidad de que se dieran ambas a la vez, y por consiguiente, de que el modelo fallara sería mayor. Si hubiera correlación negativa, la probabilidad de error sería menor. Por tanto, la condición de exclusividad de los diagnósticos es razonable.

2ª hipótesis: los hallazgos son condicionalmente independientes para cada uno de los diagnósticos.

Esta hipótesis se ha introducido al dibujar el grafo causal. Esta hipótesis será cierta si los mecanismos por los que cada enfermedad genera cada uno de los hallazgos son independientes (de la independencia causal deducimos la independencia probabilista) y si no hay ninguna otra causa (a parte de esas enfermedades) que produzca correlaciones entre los hallazgos.

A estas dos hipótesis, inherentes al método probabilista clásico, hemos añadido una tercera al suponer que la probabilidad de un falso positivo en la prueba (análisis de sangre) que detecta *A* es la misma para *b* y para *n*. Esto podría ser falso si, por ejemplo, la enfermedad *B* fuera causante de falsos positivos en mayor medida que cuando no hay ninguna enfermedad (*n*).

A continuación, se expondrán las probabilidades de cada una de las enfermedades en función de los tres hallazgos: el síntoma, la prueba genética y el análisis de sangre. Para realizar estos cálculos se supondrá que se conoce si el paciente presenta el síntoma, pero no siempre se han realizado pruebas de laboratorio. Por tanto, para cada una de estas pruebas hay tres situaciones posibles: positiva, negativa o no realizada.

Según el método probabilístico clásico, cuando tenemos información sobre la presencia o ausencia del síntoma pero no sobre el test genético (*G*) ni sobre el análisis de sangre (*L)* (filas 2ª y 3ª de la tabla siguiente) la ecuación que se debe aplicar es

$$P(d|s)=\alpha\cdot P(d)\cdot P(s|d)$$

donde *d* toma tres valores: ***a*** *,* ***b*** y ***n***; cuando se dispone de información sobre el síntoma y el test genético (filas 4ª a 6ª) la ecuación es

$$P(d|s,g)=\alpha\cdot P(d)\cdot P(s|d)\cdot P(g|d)$$

cuando se tiene información tanto sobre el síntoma como sobre el análisis de sangre, la ecuación es

$$P(d|s,l)=\alpha\cdot P(d)\cdot P(s|d)\cdot P(l|d)$$

y cuando se dispone de información del síntoma, el test genético y el análisis de sangre, la ecuación es

$$P(d|s,g,l)=\alpha\cdot P(d)\cdot P(s|d)\cdot P(g|d)\cdot P(l|d)$$

Por tanto, la probabilidad de cada diagnóstico *d* dada la evidencia **e***,* $P(d|e)$ , es la siguiente:

| **Caso** | *e* | *a* | *b* | *n* |
|---|---|---|---|---|
| 1 | $+s$ | **0,3605** | **0,5373** | **0,1022** |
| 2 | $\neg s$ | **0,0075** | **0,0029** | **0,9896** |
| 3 | $+s,+g$ | **0,1623** | **0,8358** | **0,0019** |
| 4 | $+s,\neg g$ | **0,5028** | **0,3231** | **0,1741** |
| 5 | $\neg s,+g$ | **0,1270** | **0,1696** | **0,7034** |
| 6 | $\neg s,\neg g$ | **0,0062** | **0,0010** | **0,9928** |
| 7 | $+s,+l$ | **0,9813** | **0,0157** | **0,0030** |
| 8 | $+s,\neg l$ | **0,0383** | **0,8080** | **0,1537** |
| 9 | $\neg s,+l$ | **0,4134** | **0,0017** | **0,5849** |
| 10 | $\neg s,\neg l$ | **0,0005** | **0,0029** | **0,9965** |
| 11 | $+s,+g,+l$ | **0,9474** | **0,0525** | **0,0001** |
| 12 | $+s,+g,\neg l$ | **0,0135** | **0,9842** | **0,0023** |
| 13 | $+s,\neg g,+l$ | **0,9895** | **0,0068** | **0,0037** |
| 14 | $+s,\neg g,\neg l$ | **0,0667** | **0,6065** | **0,3268** |
| 15 | $\neg s,+g,+l$ | **0,9312** | **0,0134** | **0,0554** |
| 16 | $\neg s,+g,\neg l$ | **0,0102** | **0,1923** | **0,7975** |
| 17 | $\neg s,\neg g,+l$ | **0,3662** | **0,0007** | **0,6332** |
| 18 | $\neg s,\neg g,\neg l$ | **0,0004** | **0,0010** | **0,9985** |

Para completar la tabla se han realizado todos los cálculos “a mano”. Con la intención de ilustrar cómo se han realizado, se exponen los cálculos asociados a tres de los casos que aparecen en la tabla anterior.

**Caso 1**: $\boxed{+s}$

Para resolver este caso se hace uso de la siguiente ecuación:

$$P(d|s)=\alpha\cdot P(d)\cdot P(s|d)$$

Hay que tener en cuenta que

$$\alpha=\frac{1}{\sum_{d'} P(d')\cdot P(s|d')}$$

donde $d'$ recorre todos los valores que toma la variable *D*.

En primer lugar, se calcula $\alpha$ .

$$\alpha=\frac{1}{\sum_{d'} P(d')\cdot P(s|d')}=\frac{1}{P(a)\cdot P(+s|a)+P(b)\cdot P(+s|b)+P(n)\cdot P(+s|n)}=$$

$$=\frac{1}{0{,}0239\cdot 0{,}7+0{,}0277\cdot 0{,}9+0{,}9484\cdot 0{,}005}=\frac{1}{0{,}046402}$$

Una vez calculados todos los “elementos” que aparecen en la ecuación $P(d|s)=\alpha\cdot P(d)\cdot P(s|d)$ se procede al cálculo de las distintas probabilidades.

$$P(a|+s)=\alpha\cdot P(a)\cdot P(+s|a)=\frac{1}{0{,}046402}\cdot 0{,}0239\cdot 0{,}7=0{,}3605$$

$$P(b|+s)=\alpha\cdot P(b)\cdot P(+s|b)=\frac{1}{0{,}046402}\cdot 0{,}0277\cdot 0{,}9=0{,}5373$$

$$P(n|+s)=\alpha\cdot P(n)\cdot P(+s|n)=\frac{1}{0{,}046402}\cdot 0{,}9484\cdot 0{,}005=0{,}1022$$

**Caso 4** : $\boxed{+s,\neg g}$

Para resolver este caso se hace uso de la siguiente ecuación:

$$P(d|s,g)=\alpha\cdot P(d)\cdot P(s|d)\cdot P(g|d)$$

Hay que tener en cuenta que

$$\alpha = \frac{1}{\sum_{d'} P(d') \cdot P(s, g | d')}$$

donde $d'$ recorre todos los valores que toma la variable *D*.

En primer lugar, se calcula $\alpha$ .

$$\alpha = \frac{1}{\sum_{d'} P(d') \cdot P(+s, \neg g | d')} = \frac{1}{P(a) \cdot P(+s, \neg g | a) + P(b) \cdot P(+s, \neg g | b) + P(n) \cdot P(+s, \neg g | n)} =$$

$$= \frac{1}{P(a) \cdot P(+s | a) \cdot P(\neg g | a) + P(b) \cdot P(+s | b) \cdot P(\neg g | b) + P(n) \cdot P(+s | n) \cdot P(\neg g | n)} =$$

$$= \frac{1}{0{,}0239 \cdot 0{,}7 \cdot 0{,}812 + 0{,}0277 \cdot 0{,}9 \cdot 0{,}3502 + 0{,}9484 \cdot 0{,}005 \cdot 0{,}99209} = \frac{1}{0{,}02702}$$

Una vez calculados todos los "elementos" que aparecen en la ecuación se procede al cálculo de las distintas probabilidades.

$$P(a | +s, \neg g) = \alpha \cdot P(a) \cdot P(+s | a) \cdot P(\neg g | a) = \frac{1}{0{,}02702} \cdot 0{,}0239 \cdot 0{,}7 \cdot 0{,}812 = 0{,}5028$$

$$P(b | +s, \neg g) = \alpha \cdot P(b) \cdot P(+s | b) \cdot P(\neg g | b) = \frac{1}{0{,}02702} \cdot 0{,}0277 \cdot 0{,}9 \cdot 0{,}3502 = 0{,}3231$$

Existe una forma alternativa de calcular $P(n | +s, \neg g)$ que evita hacer uso de la ecuación. Como la suma de las probabilidades asociadas a cada uno de los valores que toma la variable *D* tienen que sumar 1, entonces

$$P(n | +s, \neg g) = 1 - (P(a | +s, \neg g) + P(b | +s, \neg g)) = 1 - (0{,}5028 + 0{,}3231) = 0{,}1741$$

**Caso 17** : $\boxed{\neg s, \neg g, +l}$

Para resolver este caso se hace uso de la siguiente ecuación:

$$P(d | s, g, l) = \alpha \cdot P(d) \cdot P(s | d) \cdot P(g | d) \cdot P(l | d)$$

Hay que tener en cuenta que

$$\alpha = \frac{1}{\sum_{d'} P(d') \cdot P(s, g, l | d')}$$

donde $d'$ recorre todos los valores que toma la variable *D*.

En primer lugar, se calcula $\alpha$ .

$$\alpha = \frac{1}{P(a) \cdot P(\neg s, \neg g, +l|a) + P(b) \cdot P(\neg s, \neg g, +l|b) + P(n) \cdot P(\neg s, \neg g, +l|n)} =$$

$$= \frac{1}{P(a) \cdot P(\neg s|a) \cdot P(\neg g|a) \cdot P(+l|a) + P(b) \cdot P(\neg s|b) \cdot P(\neg g|b) \cdot P(+l|b) + P(n) \cdot P(\neg s|n) \cdot P(\neg g|n) \cdot P(+l|c)} =$$

$$= \frac{1}{0{,}0239 \cdot 0{,}3 \cdot 0{,}812 \cdot 0{,}93 + 0{,}0277 \cdot 0{,}1 \cdot 0{,}3502 \cdot 0{,}01 + 0{,}9484 \cdot 0{,}995 \cdot 0{,}99209 \cdot 0{,}01} = \frac{1}{0{,}014786}$$

Una vez calculados todos los “elementos” que aparecen en la ecuación se procede al cálculo de las distintas probabilidades.

$$P(a|\neg s, \neg g, +l) = \alpha \cdot P(a) \cdot P(\neg s|a) \cdot P(\neg g|a) \cdot P(+l|a) = \frac{1}{0{,}014786} \cdot 0{,}0239 \cdot 0{,}3 \cdot 0{,}812 \cdot 0{,}93 = 0{,}0366$$

$$P(b|\neg s, \neg g, +l) = \alpha \cdot P(b) \cdot P(\neg s|b) \cdot P(\neg g|b) \cdot P(+l|b) = \frac{1}{0{,}014786} \cdot 0{,}0277 \cdot 0{,}1 \cdot 0{,}3502 \cdot 0{,}01 = 0{,}0007$$

$$P(n|\neg s, \neg g, +l) = 1 - (P(a|\neg s, \neg g, +l) + P(b|\neg s, \neg g, +l)) = 1 - (0{,}0366 + 0{,}0007) = 0{,}6332$$

Paralelamente al desarrollo manual, el ejercicio se ha realizado con ayuda del software Elvira. A continuación, se muestran las imágenes, obtenidas en modo *Inferencia*, correspondientes al cálculo de las probabilidades condicionadas asociadas a cada uno de los casos planteados en la tabla de probabilidades condicionadas.

**Imagen en modo inferencia correspondiente al caso** $+s$

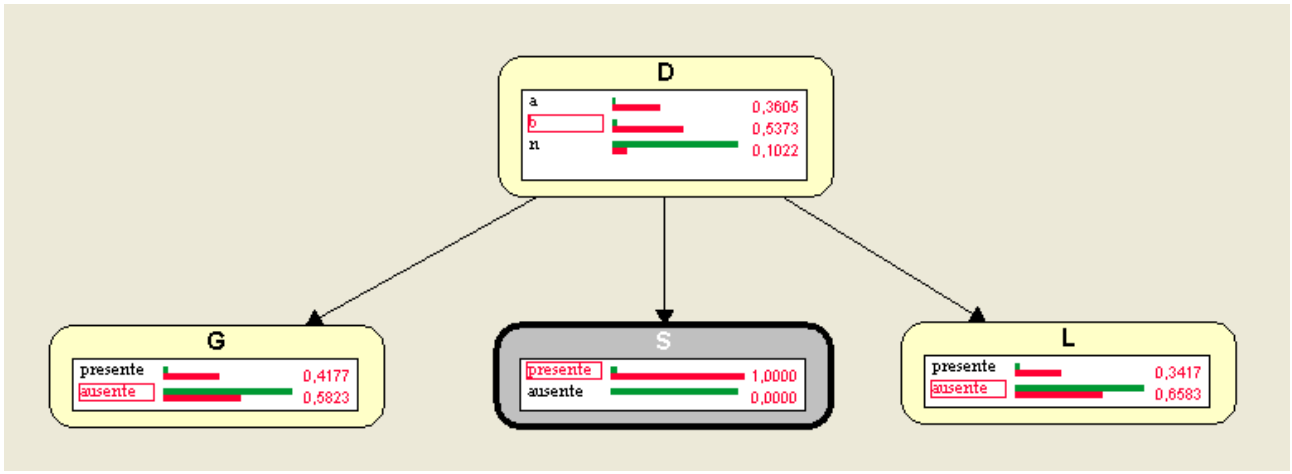



**Imagen en modo inferencia correspondiente al caso** $\neg s$

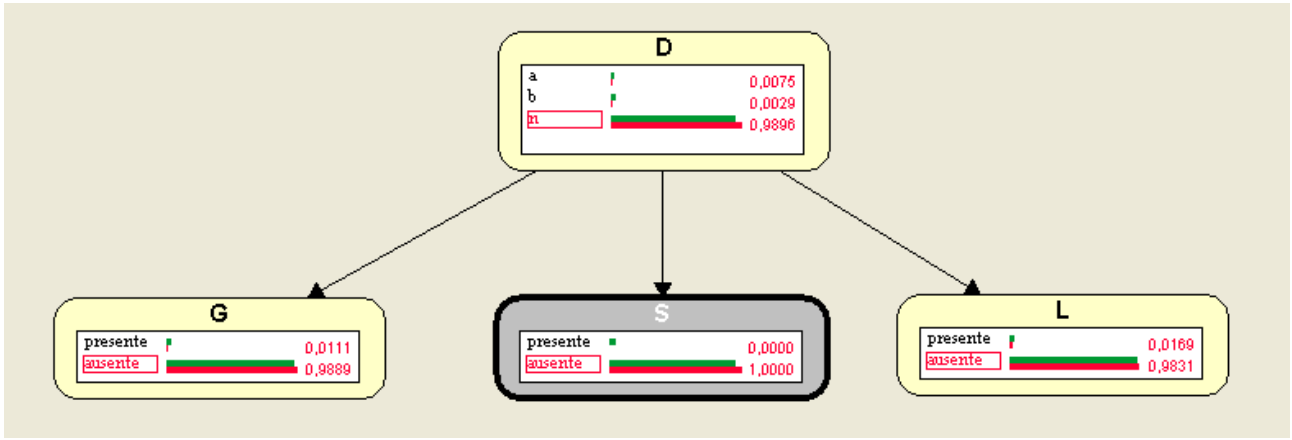



**Imagen en modo inferencia correspondiente al caso** $+s, +g$

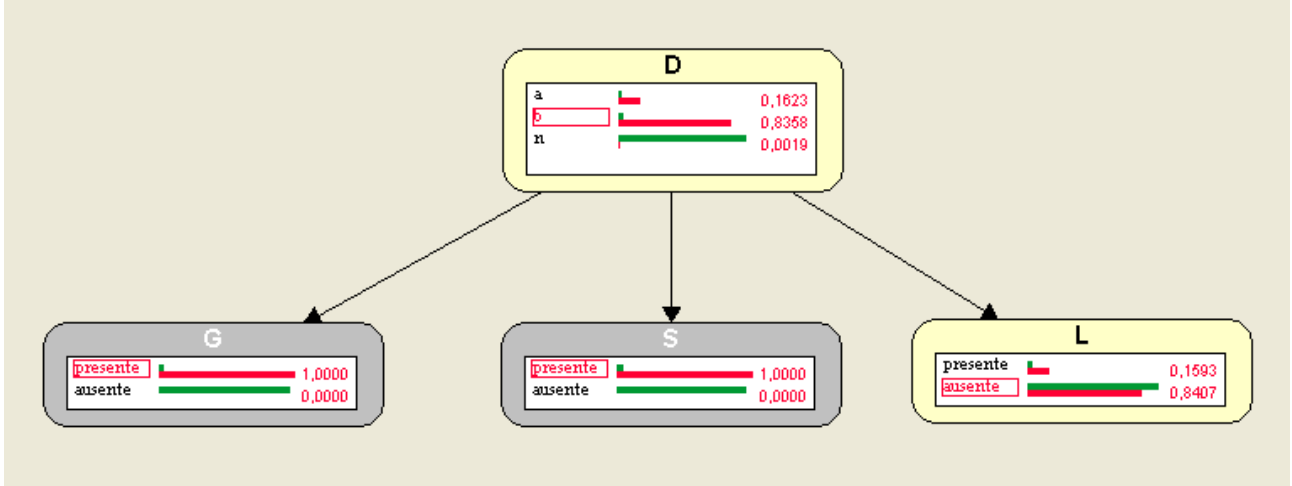


**Imagen en modo inferencia correspondiente al caso** $+s, \neg g$

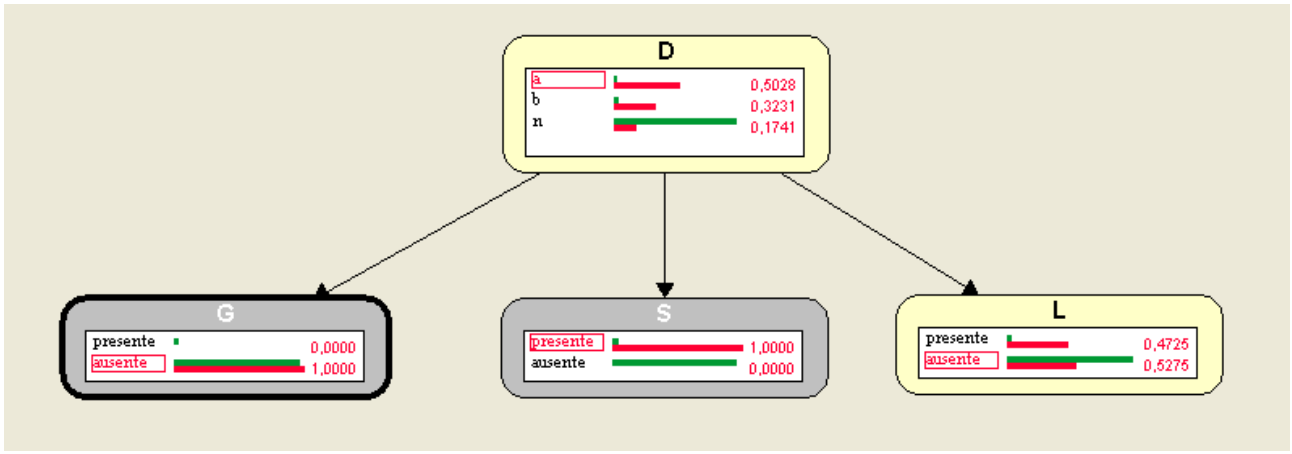



**Imagen en modo inferencia correspondiente al caso** $\neg s, +g$

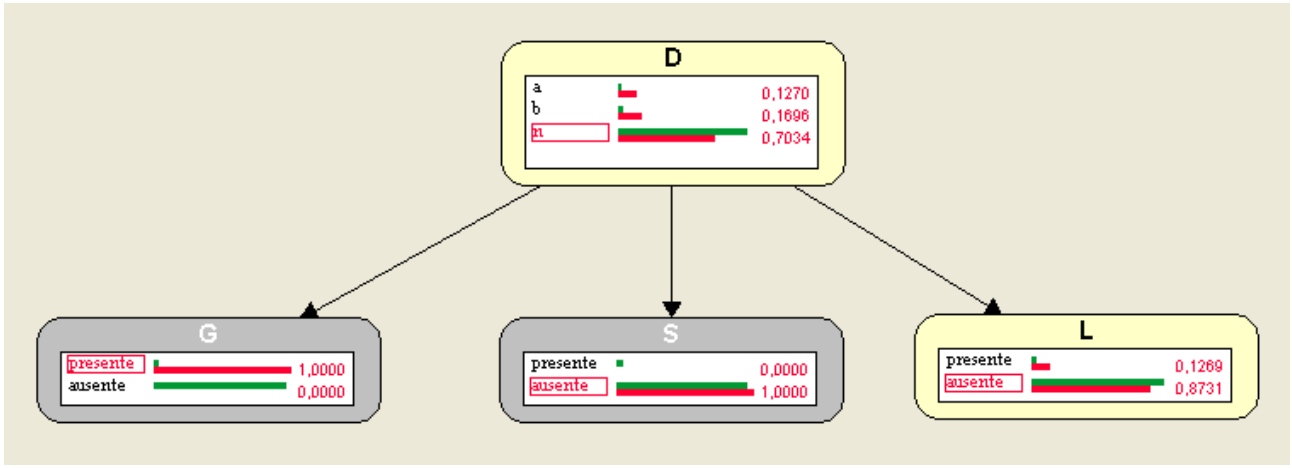



**Imagen en modo inferencia correspondiente al caso** $\neg s, \neg g$

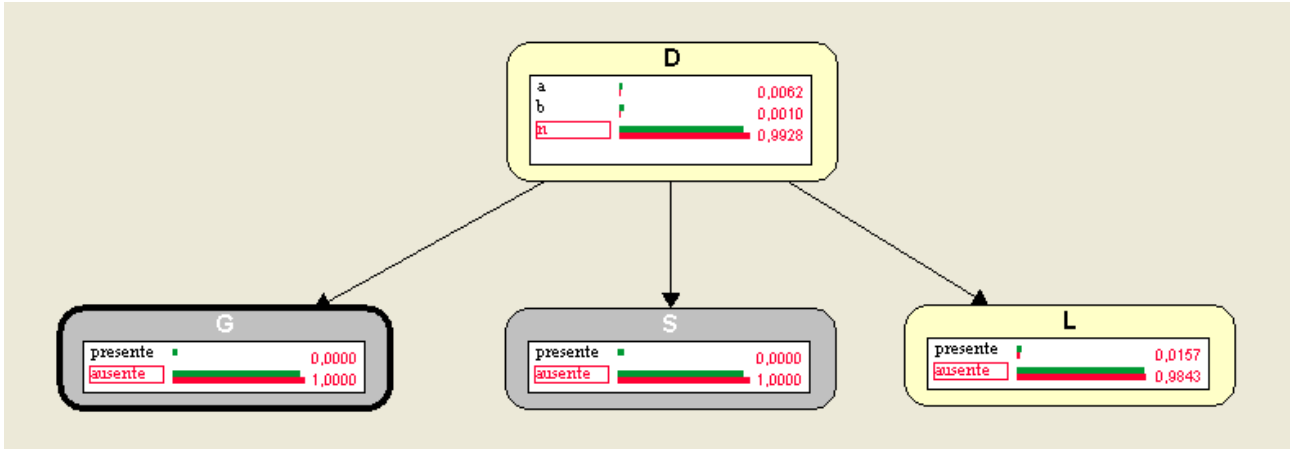


**Imagen en modo inferencia correspondiente al caso** $+s, +l$

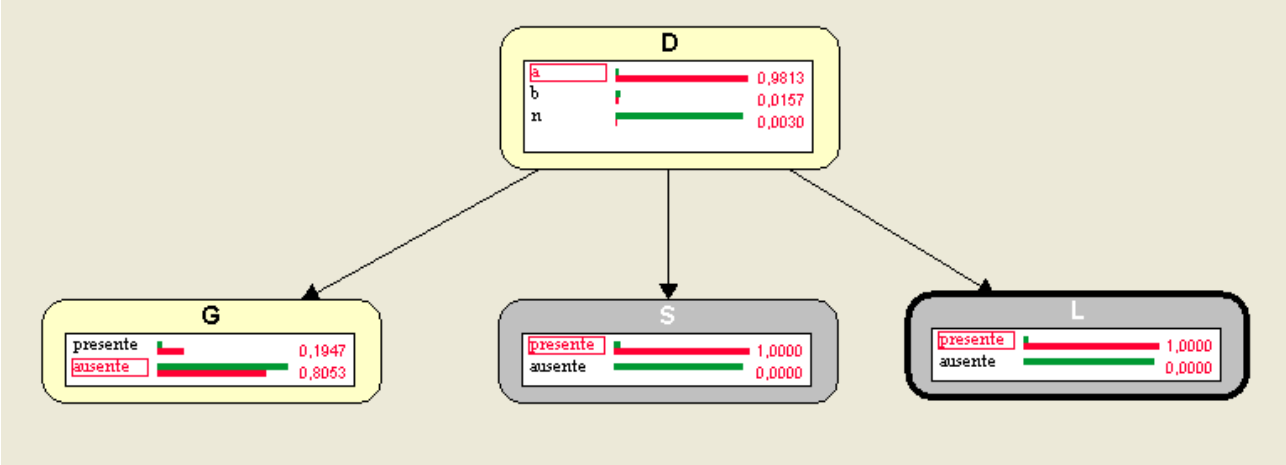



**Imagen en modo inferencia correspondiente al caso** $+s, \neg l$

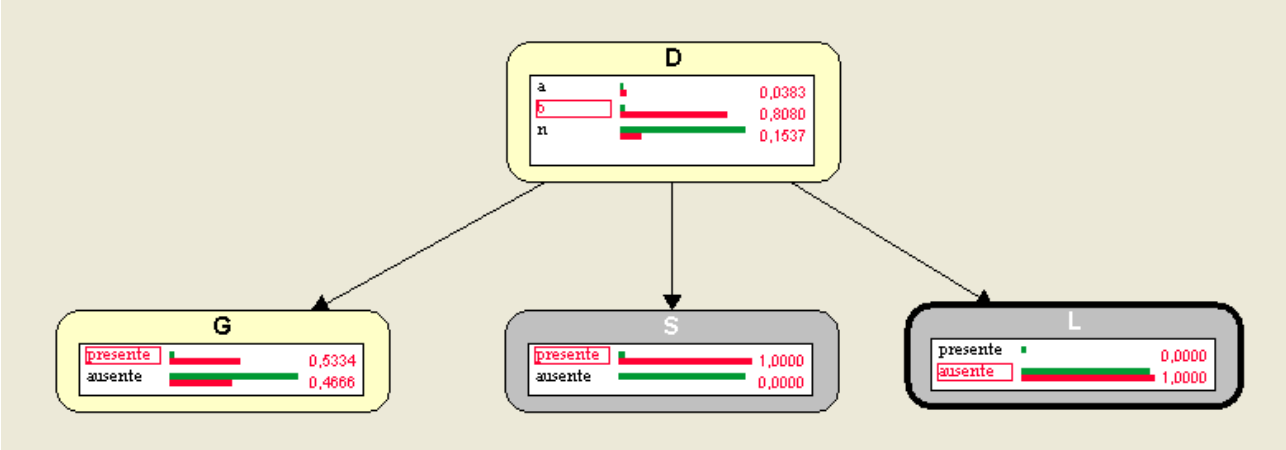



**Imagen en modo inferencia correspondiente al caso** $\neg s, +l$

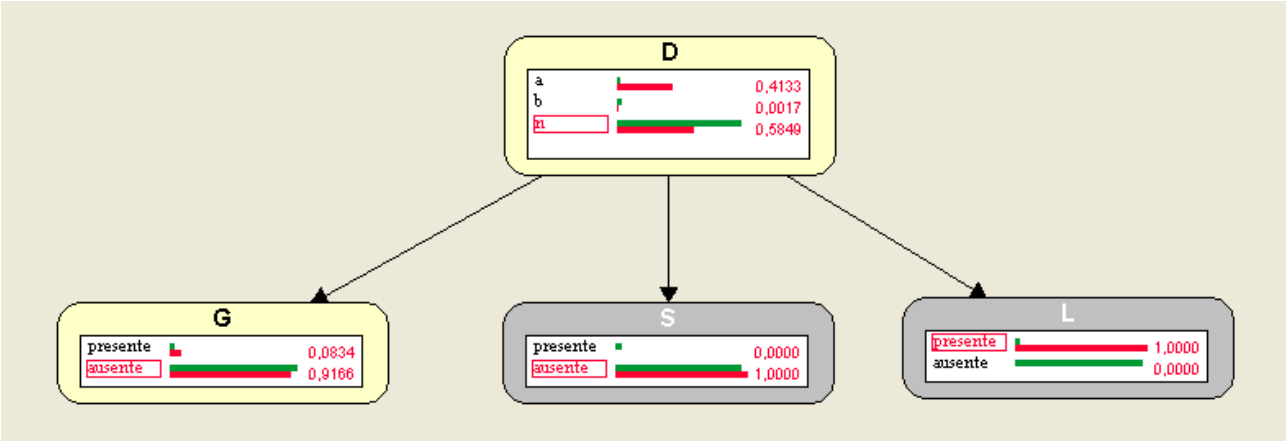


**Imagen en modo inferencia correspondiente al caso** $\boxed{\neg s, \neg l}$

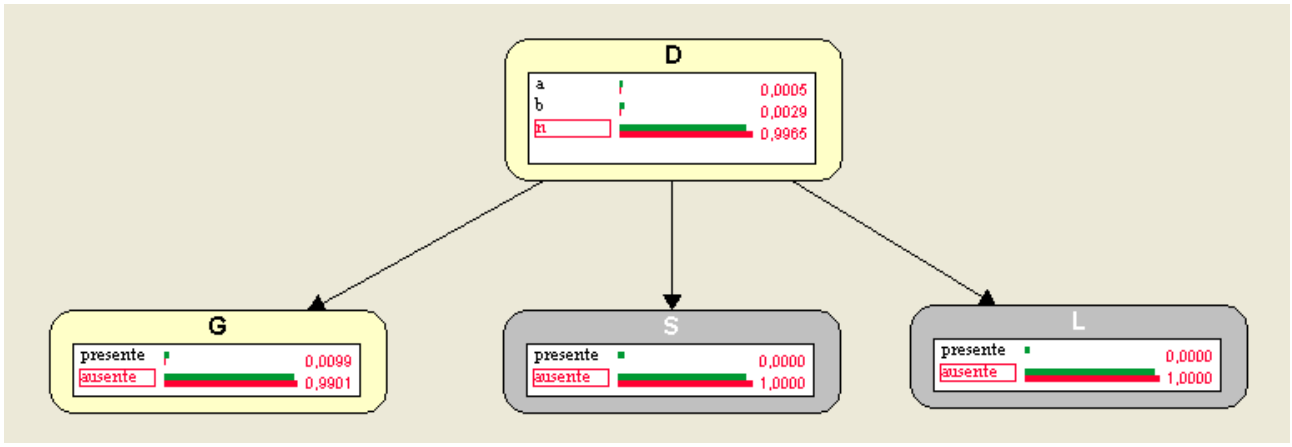



**Imagen en modo inferencia correspondiente al caso** $\boxed{+s, +g, +l}$

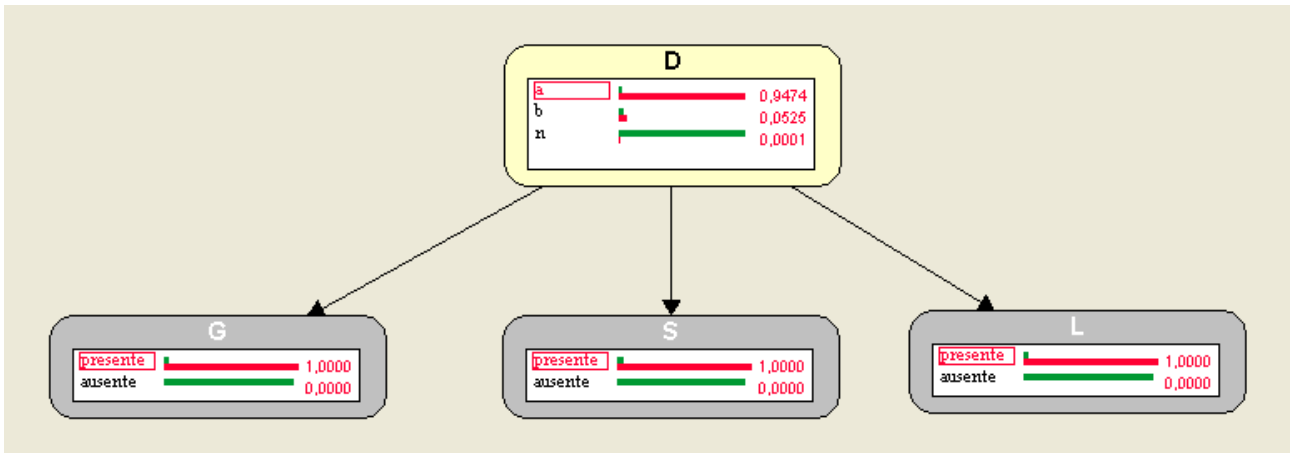



**Imagen en modo inferencia correspondiente al caso** $\boxed{+s, +g, \neg l}$

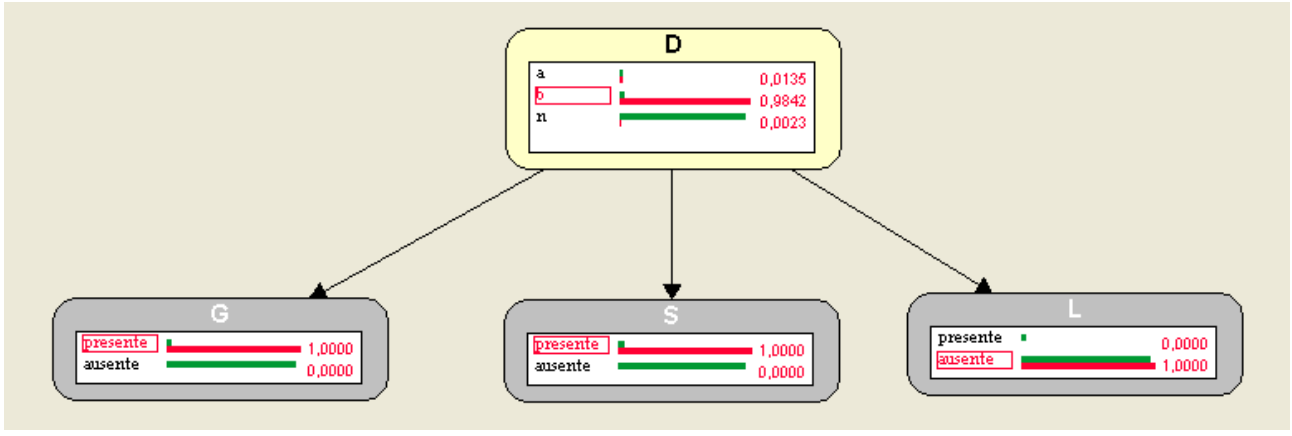


**Imagen en modo inferencia correspondiente al caso** $+s, \neg g, +l$

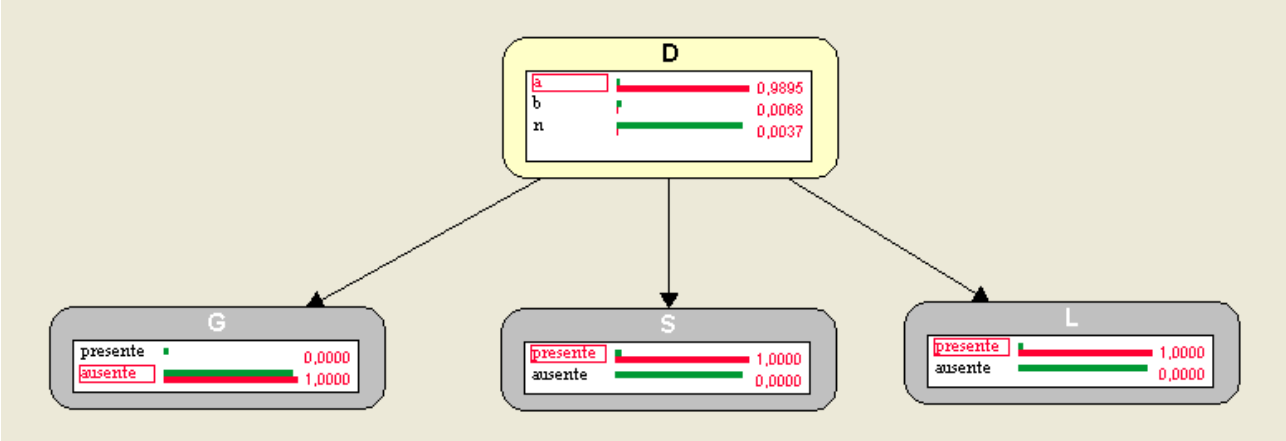



**Imagen en modo inferencia correspondiente al caso** $+s, \neg g, \neg l$

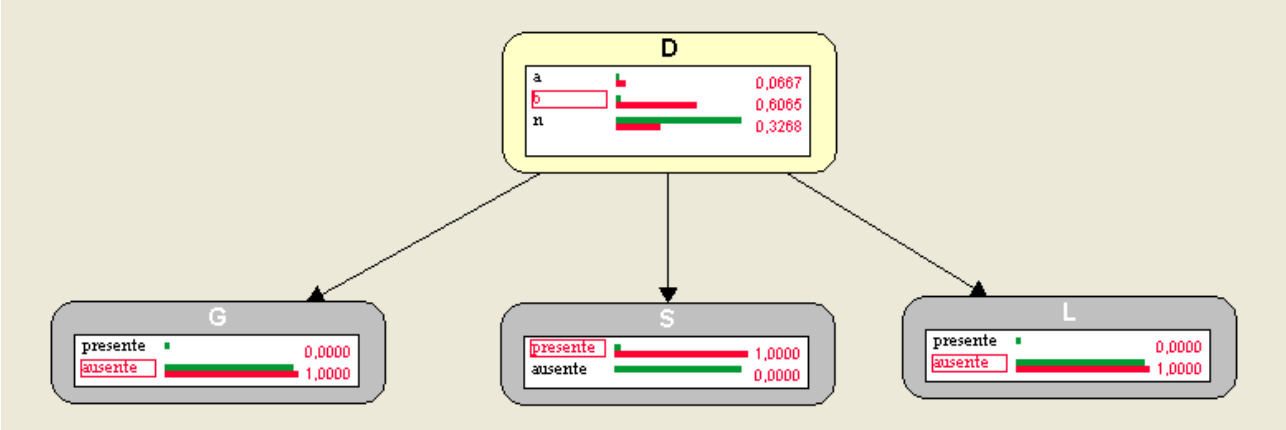



**Imagen en modo inferencia correspondiente al caso** $\neg s, +g, +l$

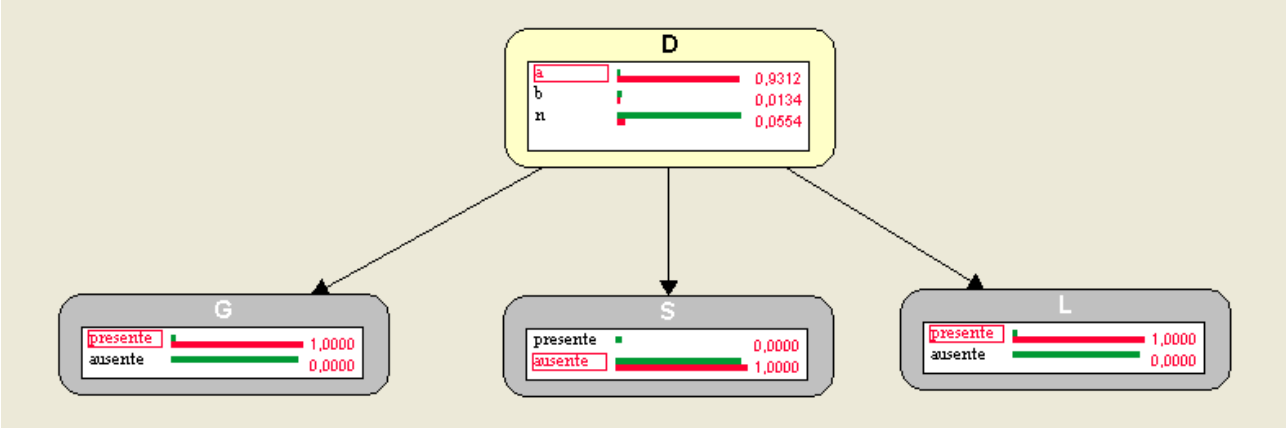


**Imagen en modo inferencia correspondiente al caso** $\boxed{\neg s, +g, \neg l}$

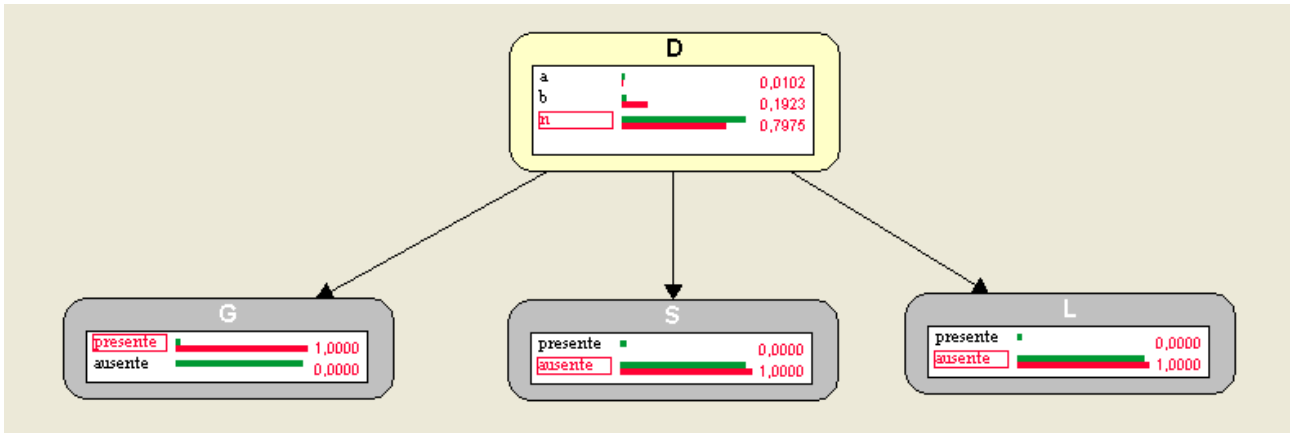



**Imagen en modo inferencia correspondiente al caso** $\boxed{\neg s, \neg g, +l}$

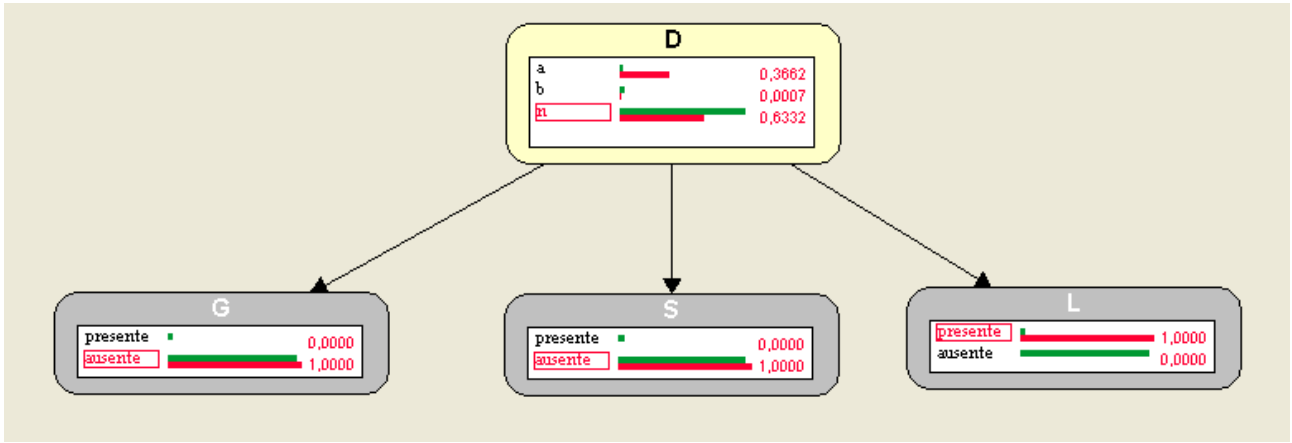
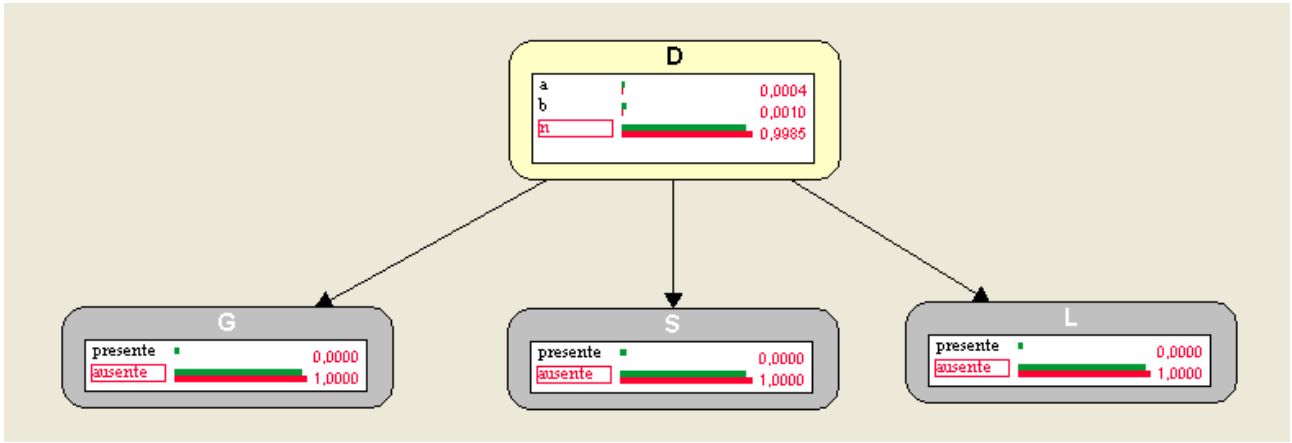



**Imagen en modo inferencia correspondiente al caso** $\boxed{\neg s, \neg g, \neg l}$



Tal y como se ha planteado este problema, no se podría resolver aplicando razones de probabilidad y de verosimilitud porque en este caso no tenemos dos hipótesis (presencia o ausencia de enfermedad) sino tres (*a*, *b* y *n*). El cálculo mediante razones de probabilidad sólo nos daría la probabilidad relativa entre dos diagnósticos, pero no la probabilidad absoluta. Por ejemplo, podríamos calcular $P(a|+s,\neg l)/P(b|+s,\neg l)$ , pero no $P(a|+s,\neg l)$ ni $P(b|+s,\neg l)$ .

Ahora se resolverá el mismo supuesto haciendo uso de una red bayesiana menos restrictiva que la planteada por el método probabilístico clásico.

Las variables que intervienen son:

- Variable ***G*** (Gen): $+g\ (presente)$ , $\neg g\ (ausente)$

- Variable ***A*** (Enfermedad A): $+a\ (presente)$ , $\neg a\ (ausente)$

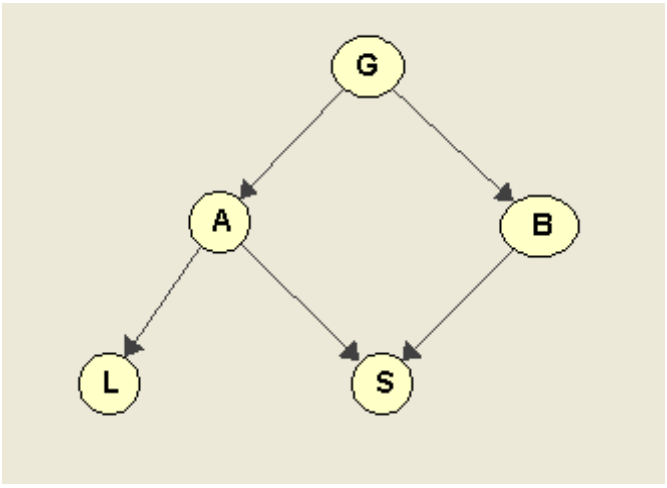
- Variable ***B*** (Enfermedad B): $+b\ (presente)$ , $\neg b\ (ausente)$

- Variable ***L*** (Análisis de sangre):, $+l\ (positivo)$ , $\neg l\ (negativo)$

- Variable ***S*** (Síntoma): $+s\ (presente)$ , $\neg s\ (ausente)$

En esta red bayesiana no hay un nodo que represente todos los diagnósticos, sino que cada enfermedad viene representada por un nodo, lo cual permitirá diagnosticar la presencia simultánea de dos enfermedades. La ausencia de enfermedad se diagnostica cuando las probabilidades de *A* y *B* son nulas o casi nulas.

El grafo causal asociado a este planteamiento es:

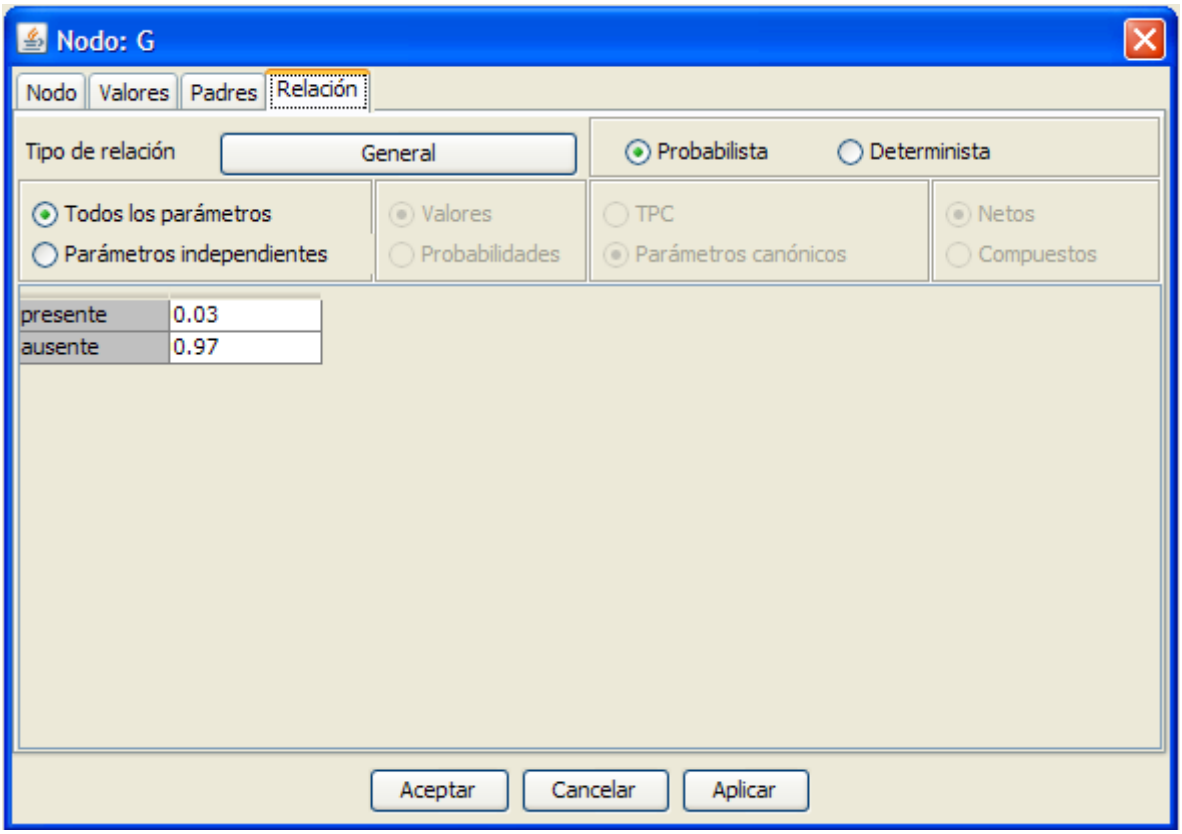


El grafo anterior “afirma” que la probabilidad de *L* no depende de *B*, mientras que el grafo asociado al método probabilístico clásico no decía si la probabilidad de *L* dependía o no de *B*. Otra diferencia sustancial es la consideración del gen *G* como causa de las dos enfermedades en esta red bayesiana, mientras en el método probabilístico clásico no se podía reflejar esta relación causal debido a que sólo se incluyen como hallazgos los efectos de una enfermedad.

A continuación, se muestran las tablas de probabilidades condicionadas acompañadas de su imagen correspondiente a través de *Elvira*.

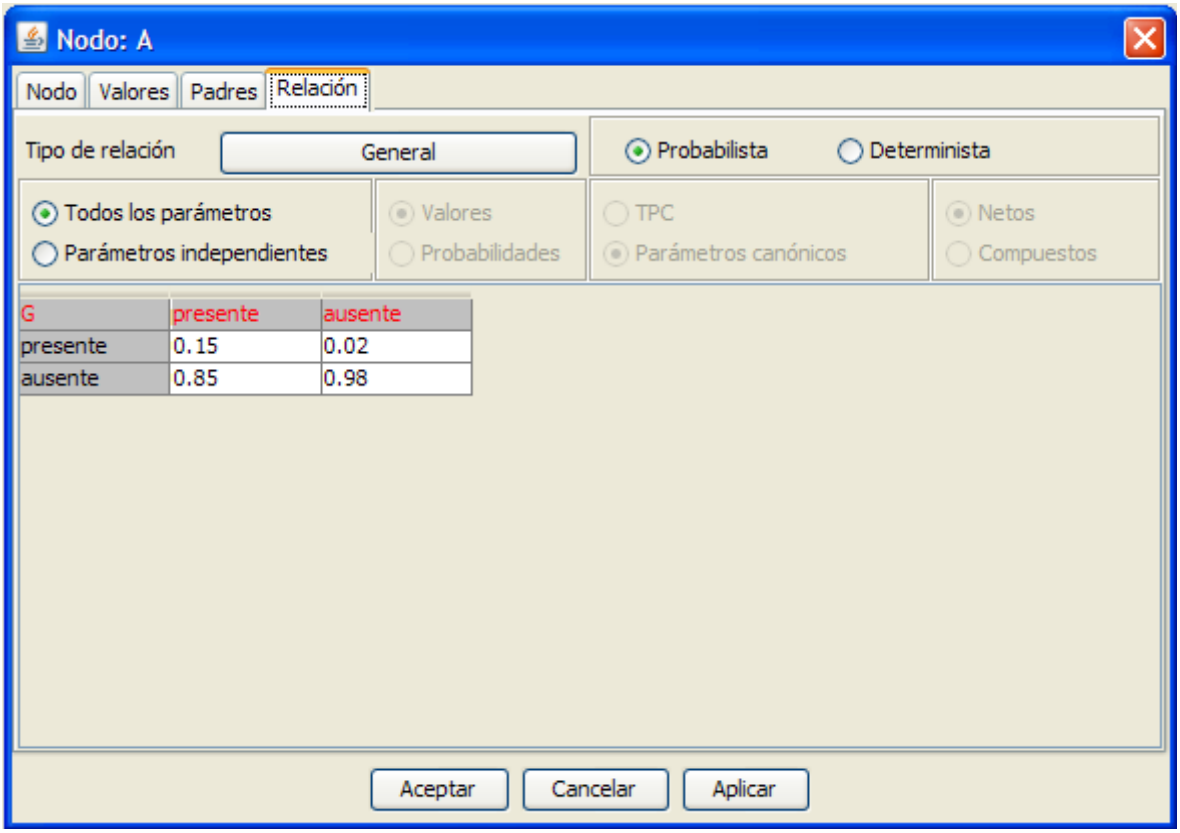
La tabla de probabilidades asociada a la variable *G* es:

| | $+g$ | $\neg g$ |
|---|---|---|
| $P(G)$ | **0,03** | **0,97** |



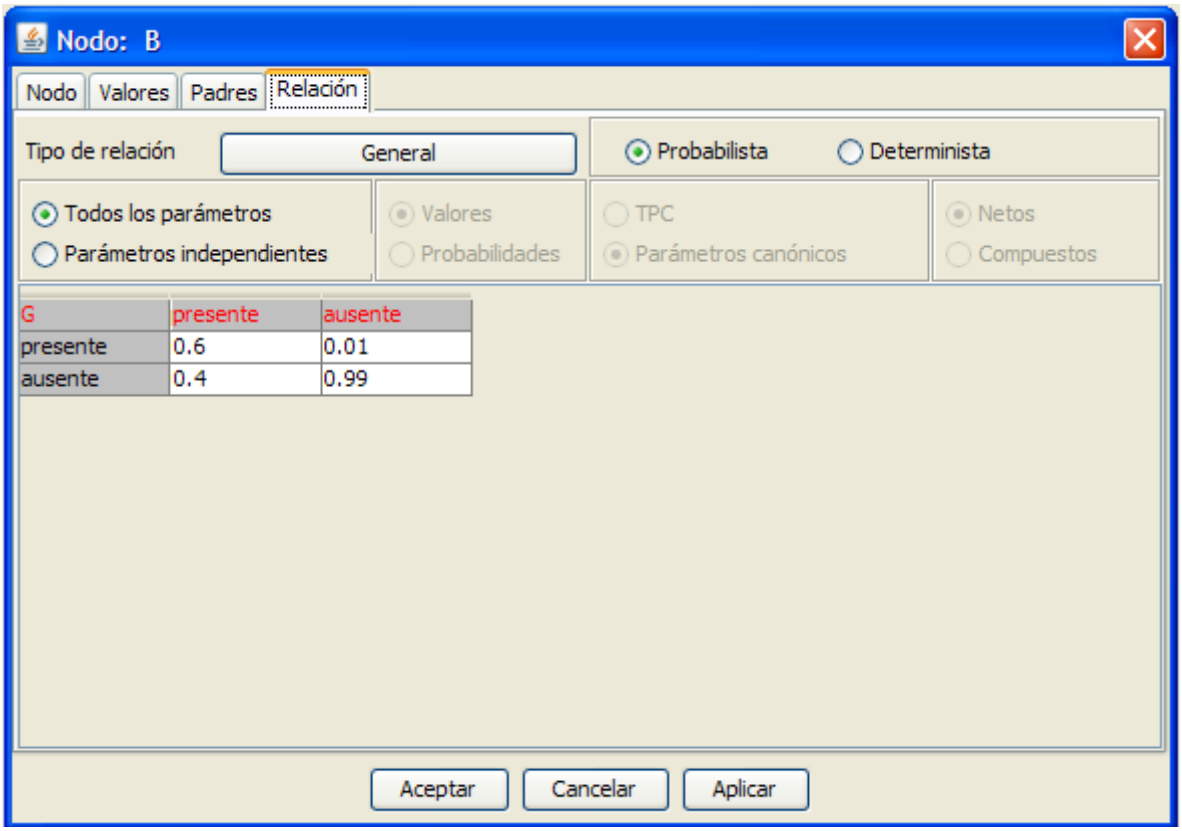
La tabla de probabilidades condicionadas asociada a la variable *A* es:

| $P(A\|G)$ | $+g$ | $\neg g$ |
|---|---|---|
| $+a$ | **0,15** | **0,02** |
| $\neg a$ | **0,85** | **0,98** |



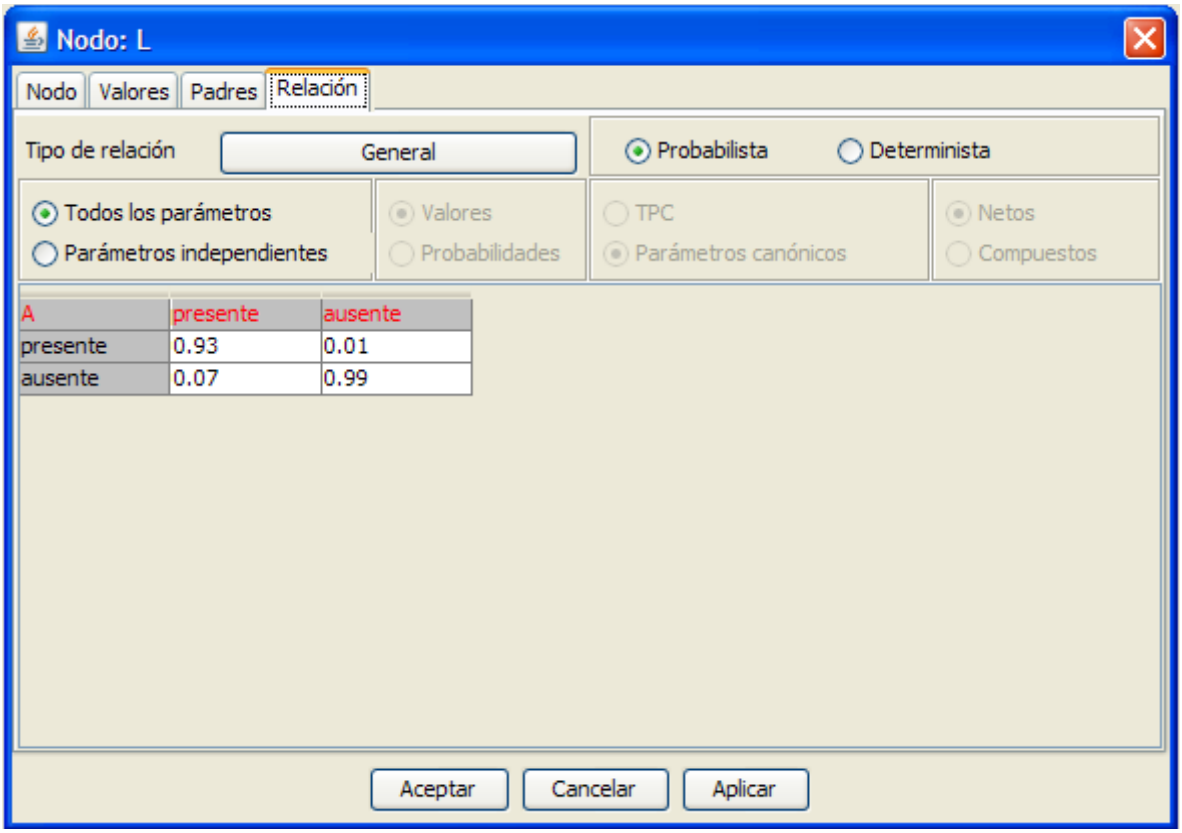

La tabla de probabilidades condicionadas asociada a la variable *B* es:

| $P(B\|G)$ | $+g$ | $\neg g$ |
|---|---|---|
| $+b$ | **0,60** | **0,01** |
| $\neg b$ | **0,40** | **0,99** |




La tabla de probabilidades condicionadas asociada a la variable *L* es:

| $P(L\|A)$ | $+a$ | $\neg a$ |
|---|---|---|
| $+l$ | **0,93** | **0,01** |
| $\neg l$ | **0,07** | **0,99** |




En el enunciado del supuesto no se dispone de suficiente información para completar la tabla $P(S|A,B)$ . Para salvar esta dificultad se supondrá que las dos enfermedades interactúan mediante una puerta OR residual al producir el síntoma *S*. La probabilidad residual, es decir, la probabilidad de que aparezca el síntoma *S* cuando las dos enfermedades están ausentes, viene dada por el enunciado del problema:

$$P(+s|\neg a,\neg b)=0{,}005$$

Si se considera esta probabilidad como la probabilidad de que el síntoma *S* sea producido por causas no explicitas en el modelo (las cuales podemos representar por *z*), la anterior probabilidad se podría reescribir del siguiente modo:

$$P(+s|\neg a,\neg b,+z)=0{,}005$$

El enunciado dice que "la enfermedad *A* produce el síntoma *S* en el 70% de los casos". Este dato puede interpretarse como la probabilidad de que *A* produzca *S* cuando las causas no explicitas en el modelo están ausentes:

$$C_1=P(+s|+a,\neg b,\neg z)=0{,}70$$

El enunciado dice también que “la enfermedad *B* produce el síntoma *S* en el 90% de los casos”. Este dato puede interpretarse como la probabilidad de que *B* produzca *S* cuando las causas no explicitas en el modelo están ausentes:

$$C_2 = P(+s|\neg a, +b, \neg z) = 0{,}90$$

Y la siguiente probabilidad ya se expuso con anterioridad.

$$C_3 = P(+s|\neg a, \neg b, +z) = 0{,}005$$

Por tanto, cuando sólo se conoce que *A* está presente y *B* está ausente (no se sabe si alguna de las causas implícitas está presente o no), la probabilidad de *S* es:

$$P(+s|+a, \neg b) = C_1 + (1 - C_1) \cdot C_3 = 0{,}70 + 0{,}30 \cdot 0{,}005 = 0{,}70 + 0{,}00150 = 0{,}70150$$

Por la misma razón:

$$P(+s|\neg a, +b) = C_2 + (1 - C_2) \cdot C_3 = 0{,}90 + 0{,}10 \cdot 0{,}005 = 0{,}90 + 0{,}00050 = 0{,}90050$$

$$P(+s|+a, +b) = 1 - (1 - C_1) \cdot (1 - C_2) \cdot (1 - C_3) = 1 - (0{,}3) \cdot (0{,}10) \cdot 0{,}995 = 1 - 0{,}02985 = 0{,}97015$$

No hay que olvidar que el enunciado proporciona la siguiente probabilidad:

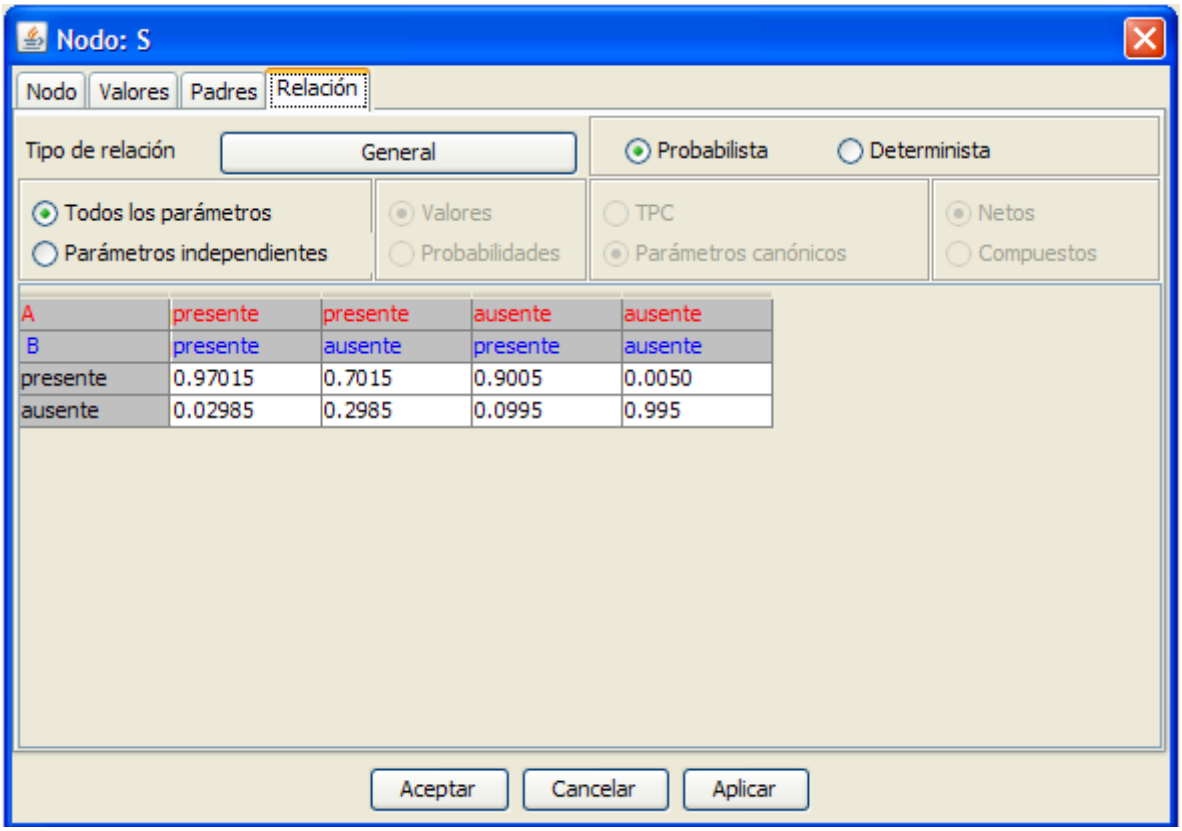
$$P(+s|\neg a, \neg b) = 0{,}005$$

A partir de las 4 probabilidades anteriores se calculan las otras 4 probabilidades mediante la expresión:

$$P(\neg s|a, b) = 1 - P(+s|a, b)$$

Por tanto, la tabla de probabilidades condicionadas asociada a la variable *S* es:

| $P(S\|A, B)$ | $+a$ | | $\neg a$ | |
|---|---|---|---|---|
| | $+b$ | $\neg b$ | $+b$ | $\neg b$ |
| $+s$ | **0,97015** | **0,7015** | **0,9005** | **0,005** |
| $\neg s$ | **0,02985** | **0,2985** | **0,0995** | **0,995** |




En la realización de este supuesto se están considerando una serie de hipótesis que conviene analizar si resultan razonables.

La ausencia de enlaces entre *A* y *B* refleja la hipótesis de que estas dos enfermedades son independientes a priori. Esta hipótesis es razonable si no hay mecanismos causales conocidos entre ellas (es decir, si ninguna enfermedad es causa de la otra) ni causas comunes (por ejemplo, un factor de riesgo en común) que no se haya detectado; en este caso se ha detectado un factor de riesgo común, el gen G, que ha sido incluido en el grafo causal trazando los enlaces correspondientes. También se vigilaría la correlación que hay entre ellos. La hipótesis de que el hallazgo *L* es independiente de *A* y *B* ya se abordó en el caso anterior.

La diferencia principal respecto del caso probabilístico clásico es que esta red bayesiana no necesita suponer que los diagnósticos son exclusivos, ni tampoco que los hallazgos sean condicionalmente independientes dado cada diagnóstico.

La probabilidad conjunta de la red bayesiana viene dada por:

$$P(g,a,b,l,s)=P(g)\cdot P(a|g)\cdot P(b|g)\cdot P(l|a)\cdot P(s|a,b)$$

A partir de ésta se pueden calcular todas las probabilidades marginales y condicionadas.

En la tabla siguiente se indica cuál es la probabilidad de cada una de las enfermedades en función de los hallazgos *S* (síntoma) y *L* (análisis de sangre) de nuestro problema, suponiendo que siempre se conoce si el paciente presenta o no el síntoma, pero no siempre se ha realizado el análisis de sangre.

| $e$ | $P(a\|e)$ | $P(b\|e)$ |
|---|---|---|
| $+s$ | **0,3930** | **0,5633** |
| $\neg s$ | **0,0067** | **0,0027** |
| $+s,+l$ | **0,9837** | **0,1709** |
| $+s,\neg l$ | **0,0438** | **0,7954** |
| $\neg s,+l$ | **0,3838** | **0,0068** |
| $\neg s,\neg l$ | **0,0005** | **0,0026** |

A continuación, se muestran de manera detallada los cálculos realizados para la obtención de dos de las probabilidades expresadas en la tabla anterior.

*Situación 1*. Para un paciente que se presenta con el síntoma *S* y da positivo en la prueba *L* (análisis de sangre), se calcula la probabilidad de que padezca *A*.

$$P(+a|+s,+l)=\frac{P(+a,+s,+l)}{P(+s,+l)}$$

$$P(+a,+s,+l)=\sum_g\sum_b P(g,+a,b,+l,+s)=\sum_g\sum_b P(g)\cdot P(+a|g)\cdot P(b|g)\cdot P(+l|+a)\cdot P(+s|+a,b)=$$

$$=P(+g)\cdot P(+a|+g)\cdot P(+b|+g)\cdot P(+l|+a)\cdot P(+s|+a,+b)+P(+g)\cdot P(+a|+g)\cdot P(\neg b|+g)\cdot P(+l|+a)\cdot$$

$$\cdot P(+s|+a,\neg b)+P(\neg g)\cdot P(+a|\neg g)\cdot P(+b|\neg g)\cdot P(+l|+a)\cdot P(+s|+a,+b)+P(\neg g)\cdot P(+a|\neg g)\cdot P(\neg b|\neg g)\cdot$$

$$\cdot P(+l|+a)\cdot P(+s|+a,b)=0{,}03\cdot 0{,}15\cdot 0{,}6\cdot 0{,}93\cdot 0{,}97015+0{,}03\cdot 0{,}15\cdot 0{,}4\cdot 0{,}93\cdot 0{,}7015+0{,}97\cdot 0{,}02\cdot 0{,}01\cdot$$

$$\cdot 0{,}93\cdot 0{,}97015+0{,}97\cdot 0{,}02\cdot 0{,}99\cdot 0{,}93\cdot 0{,}7015=0{,}01631529$$

Como $P(+s,+l)=P(+a,+s,+l)+P(\neg a,+s,+l)$ , todavía queda por calcular $P(\neg a,+s,+l)$ .

$$P(\neg a,+s,+l)=\sum_g\sum_b P(g,\neg a,b,+l,+s)=\sum_g\sum_b P(g)\cdot P(\neg a|g)\cdot P(b|g)\cdot P(+l|\neg a)\cdot P(+s|\neg a,b)=$$

$$=P(+g)\cdot P(\neg a|+g)\cdot P(+b|+g)\cdot P(+l|\neg a)\cdot P(+s|\neg a,+b)+P(+g)\cdot P(\neg a|+g)\cdot P(\neg b|+g)\cdot P(+l|\neg a)\cdot$$

$$\cdot P(+s|\neg a,\neg b)+P(\neg g)\cdot P(\neg a|\neg g)\cdot P(+b|\neg g)\cdot P(+l|\neg a)\cdot P(+s|\neg a,+b)+P(\neg g)\cdot P(\neg a|\neg g)\cdot P(\neg b|\neg g)\cdot$$

$$\cdot P(+l|\neg a)\cdot P(+s|\neg a,b)=0{,}03\cdot 0{,}85\cdot 0{,}6\cdot 0{,}01\cdot 0{,}9005+0{,}03\cdot 0{,}85\cdot 0{,}4\cdot 0{,}01\cdot 0{,}005+0{,}97\cdot 0{,}98\cdot 0{,}01\cdot$$

$$\cdot 0{,}01\cdot 0{,}9005+0{,}97\cdot 0{,}98\cdot 0{,}99\cdot 0{,}01\cdot 0{,}005=0{,}00027094273$$

$$P(+s,+l)=P(+a,+s,+l)+P(\neg a,+s,+l)=0{,}01631529+0{,}00027094273=0{,}01658623$$

$$\boxed{P(+a|+s,+l)=\frac{P(+a,+s,+l)}{P(+s,+l)}=\frac{0{,}01631529}{0{,}01658623}=0{,}983664762\approx 0{,}9837}$$

*Situación 2*. Para un paciente que no presenta el síntoma *S* y da positivo en la prueba *L* (análisis de sangre), se calcula la probabilidad de que padezca *B*.

$$P(+b|\neg s,+l)=\frac{P(+b,\neg s,+l)}{P(\neg s,+l)}$$

$$P(+b,\neg s,+l)=\sum_g\sum_a P(g,a,+b,+l,\neg s)=\sum_g\sum_a P(g)\cdot P(a|g)\cdot P(+b|g)\cdot P(+l|a)\cdot P(\neg s|a,+b)=$$

$$=0{,}03\cdot 0{,}15\cdot 0{,}60\cdot 0{,}93\cdot 0{,}02985+0{,}03\cdot 0{,}85\cdot 0{,}60\cdot 0{,}01\cdot 0{,}0995+0{,}97\cdot 0{,}02\cdot 0{,}01\cdot 0{,}93\cdot 0{,}02985+$$

$$+0{,}97\cdot 0{,}98\cdot 0{,}01\cdot 0{,}01\cdot 0{,}0995=0{,}000105020857$$

Como $P(\neg s,+l)=P(+b,\neg s,+l)+P(\neg b,\neg s,+l)$ , todavía queda por calcular $P(\neg b,\neg s,+l)$ .

$$P(\neg b,\neg s,+l)=\sum_g\sum_a P(g,a,\neg b,+l,\neg s)=\sum_g\sum_a P(g)\cdot P(a|g)\cdot P(\neg b|g)\cdot P(+l|a)\cdot P(\neg s|a,\neg b)=$$

$$=0{,}03\cdot 0{,}15\cdot 0{,}40\cdot 0{,}93\cdot 0{,}2985+0{,}03\cdot 0{,}85\cdot 0{,}40\cdot 0{,}01\cdot 0{,}995+0{,}97\cdot 0{,}02\cdot 0{,}99\cdot 0{,}93\cdot 0{,}2985+$$

$$=0{,}97\cdot 0{,}98\cdot 0{,}99\cdot 0{,}01\cdot 0{,}995=0{,}015296745$$

$$P(\neg s,+l)=P(+b,\neg s,+l)+P(\neg b,\neg s,+l)=0{,}000105020857+0{,}015296745=0{,}015401766$$

$$P(+b|\neg s,+l)=\frac{P(+b,\neg s,+l)}{P(\neg s,+l)}=\frac{0{,}0{,}000105020857}{0{,}0{,}015401766}=0{,}006818754226\approx 0{,}0068$$

A continuación, se muestran las imágenes, obtenidas mediante el modo *Inferencia* de *Elvira*, correspondientes al cálculo de las probabilidades condicionadas de distintos casos, entre los que se encuentran los de la tabla anterior.

**Imagen en modo inferencia correspondiente al caso** $+s$

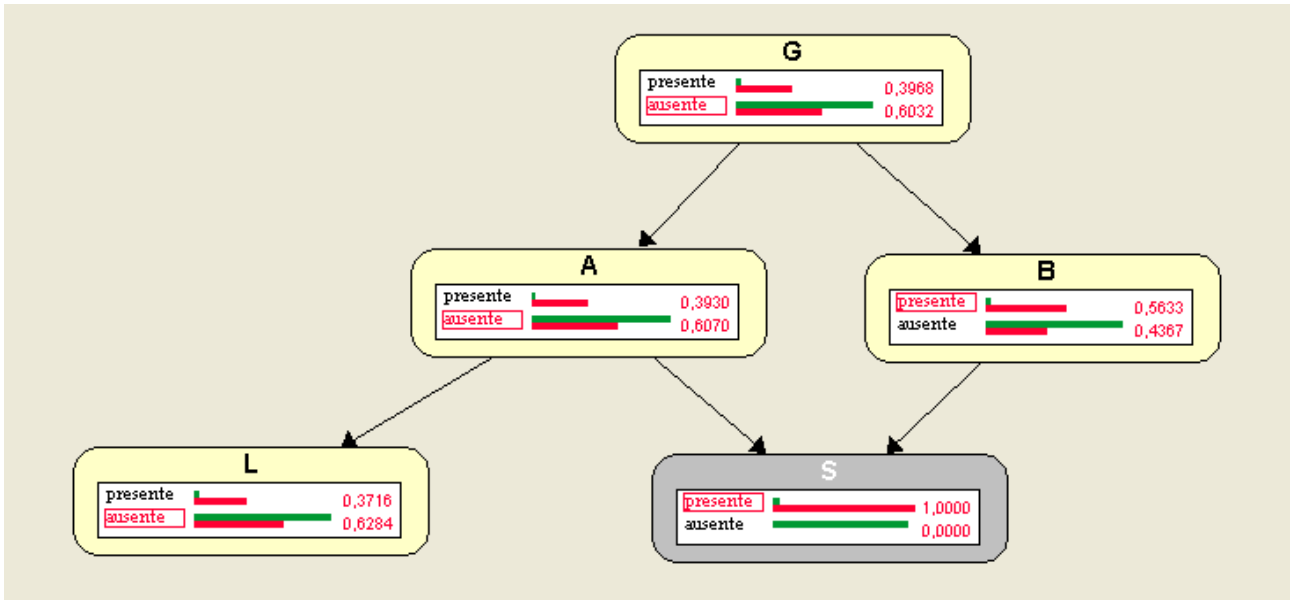



**Imagen en modo inferencia correspondiente al caso** $\neg s$

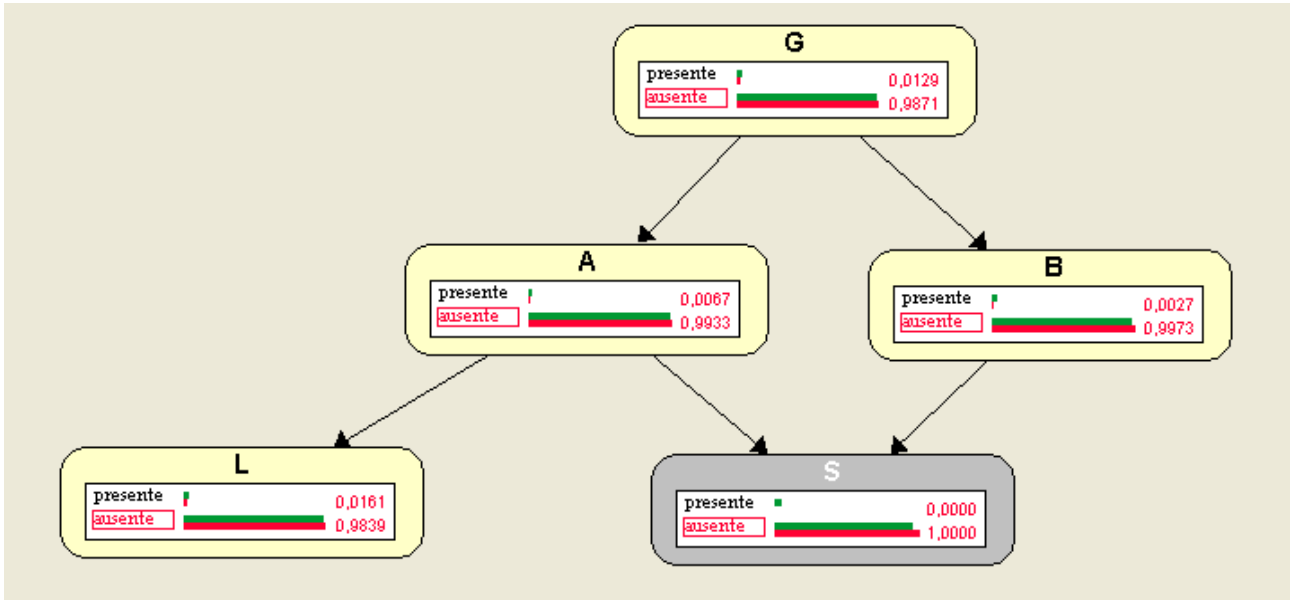


**Imagen en modo inferencia correspondiente al caso** $+s, +g$

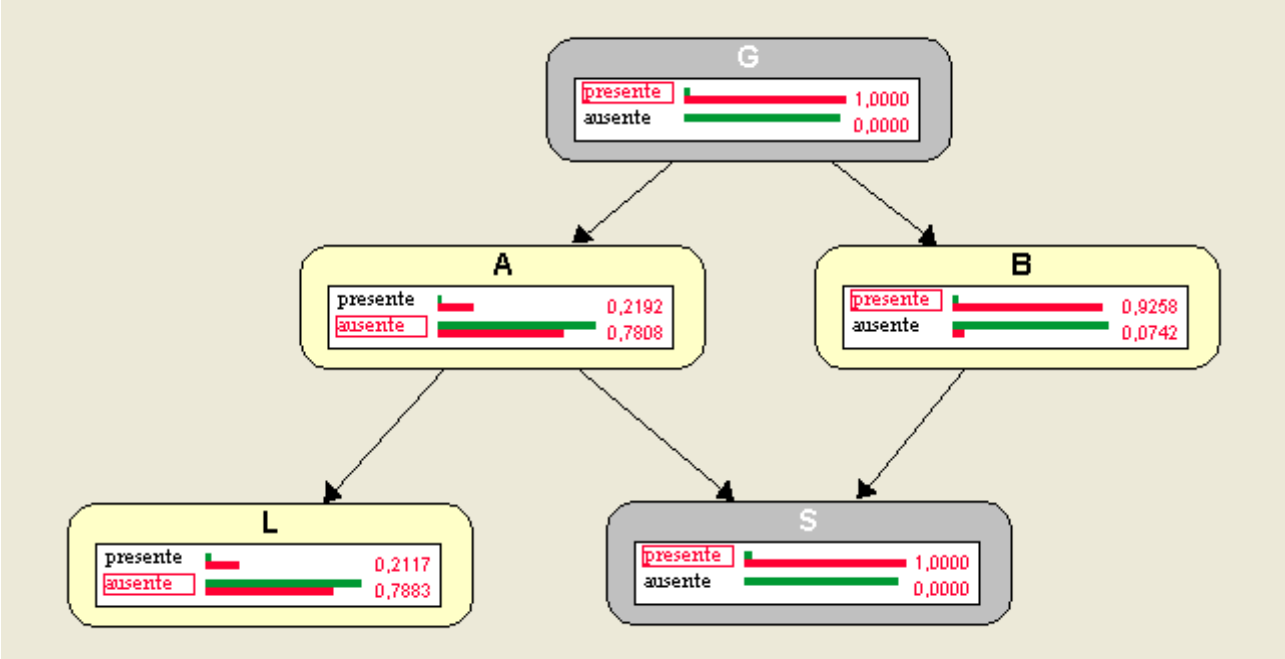



**Imagen en modo inferencia correspondiente al caso** $+s, \neg g$

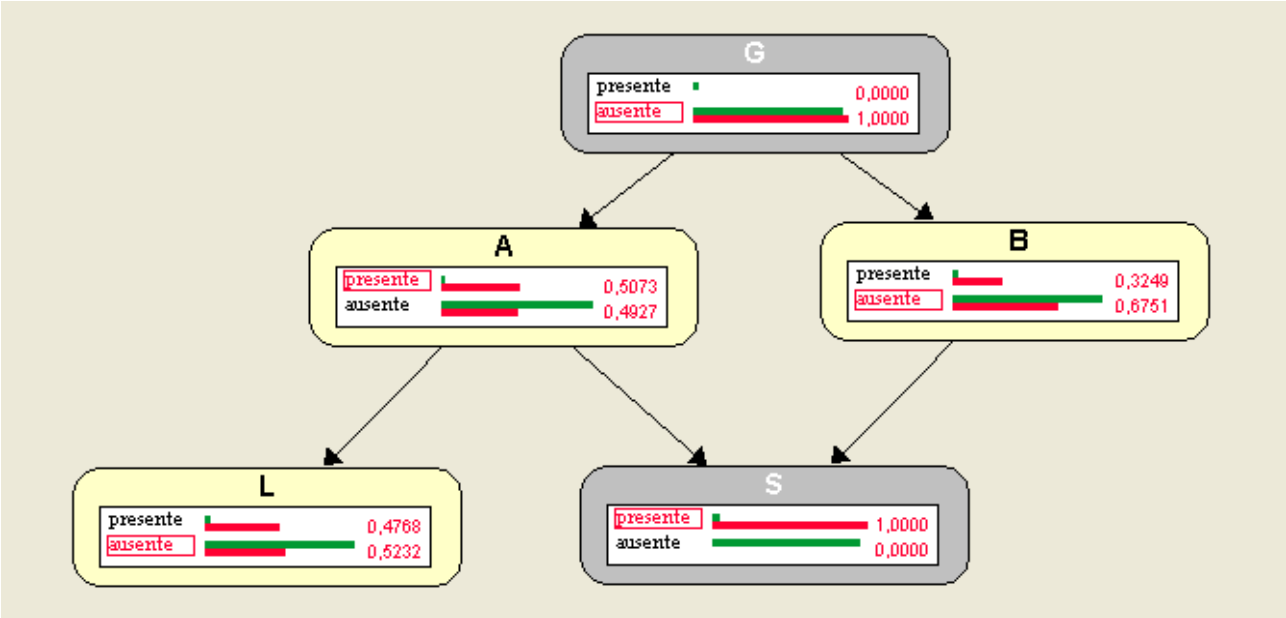


**Imagen en modo inferencia correspondiente al caso** $\neg s, +g$

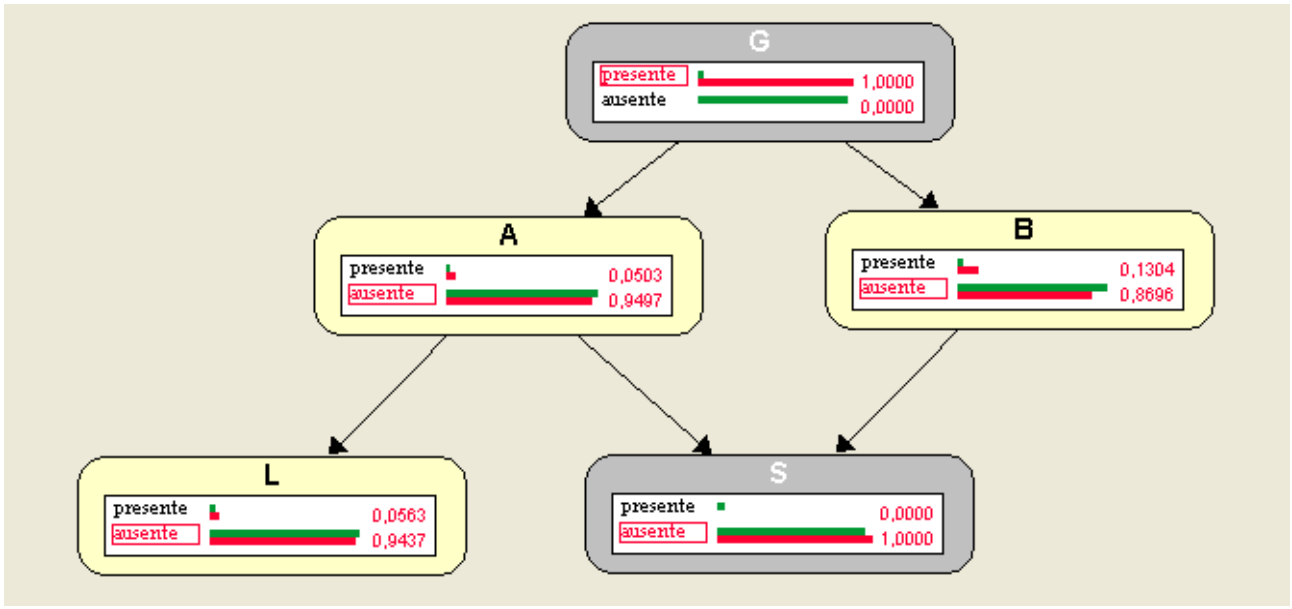



**Imagen en modo inferencia correspondiente al caso** $\neg s, \neg g$

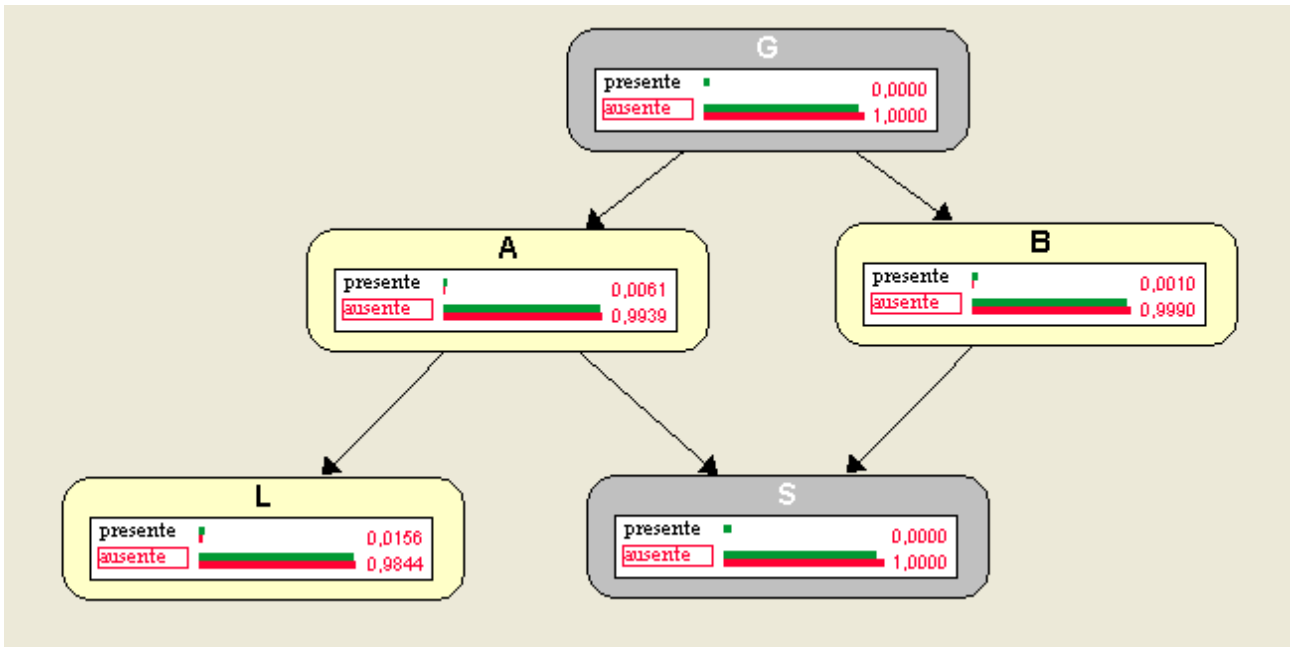


**Imagen en modo inferencia correspondiente al caso** $+s, +l$

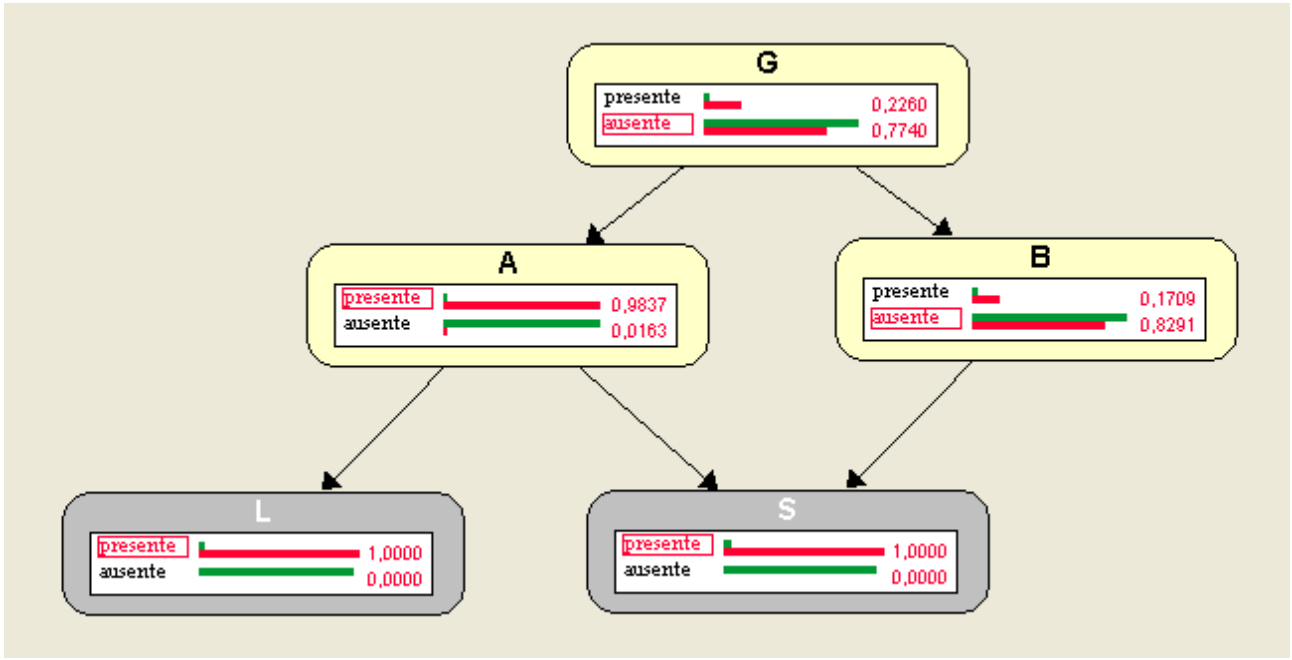



**Imagen en modo inferencia correspondiente al caso** $+s, \neg l$

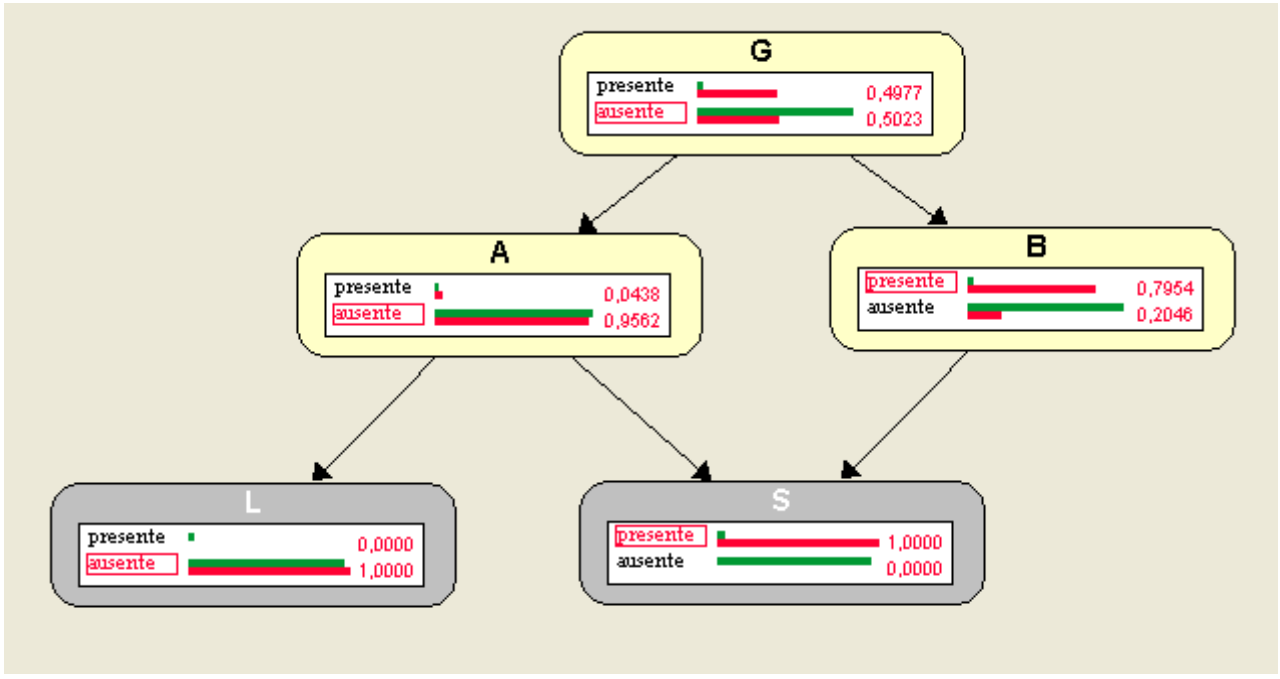


**Imagen en modo inferencia correspondiente al caso** $\boxed{\neg s, +l}$

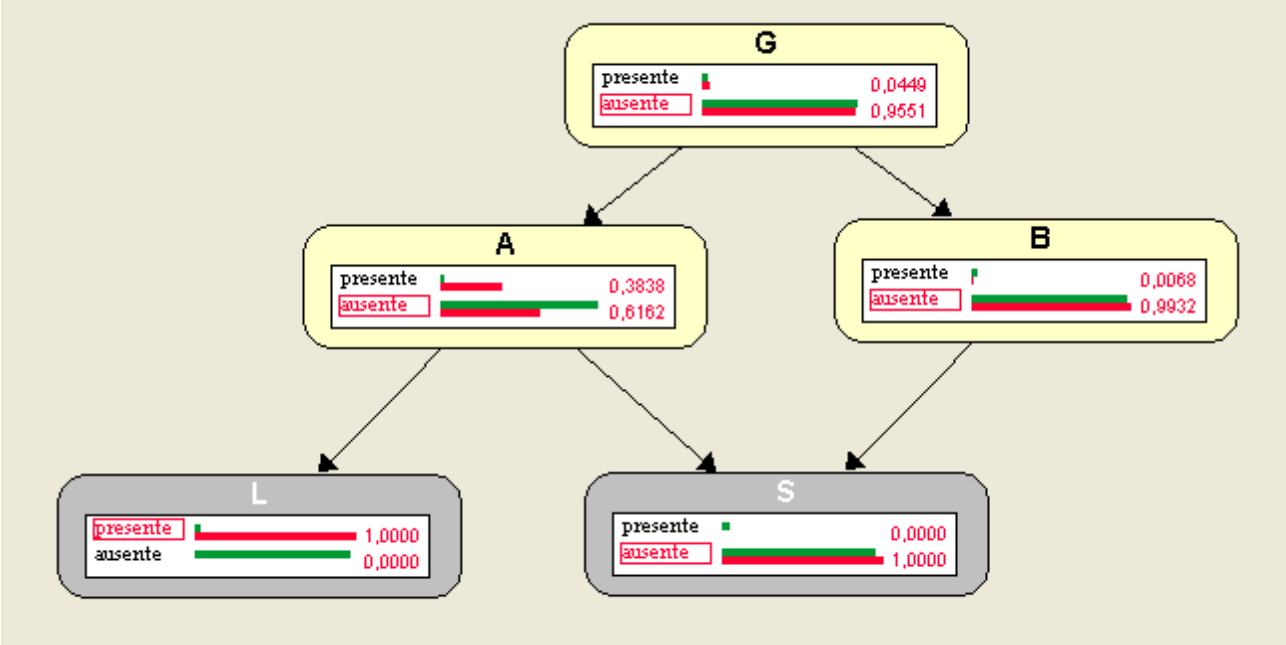
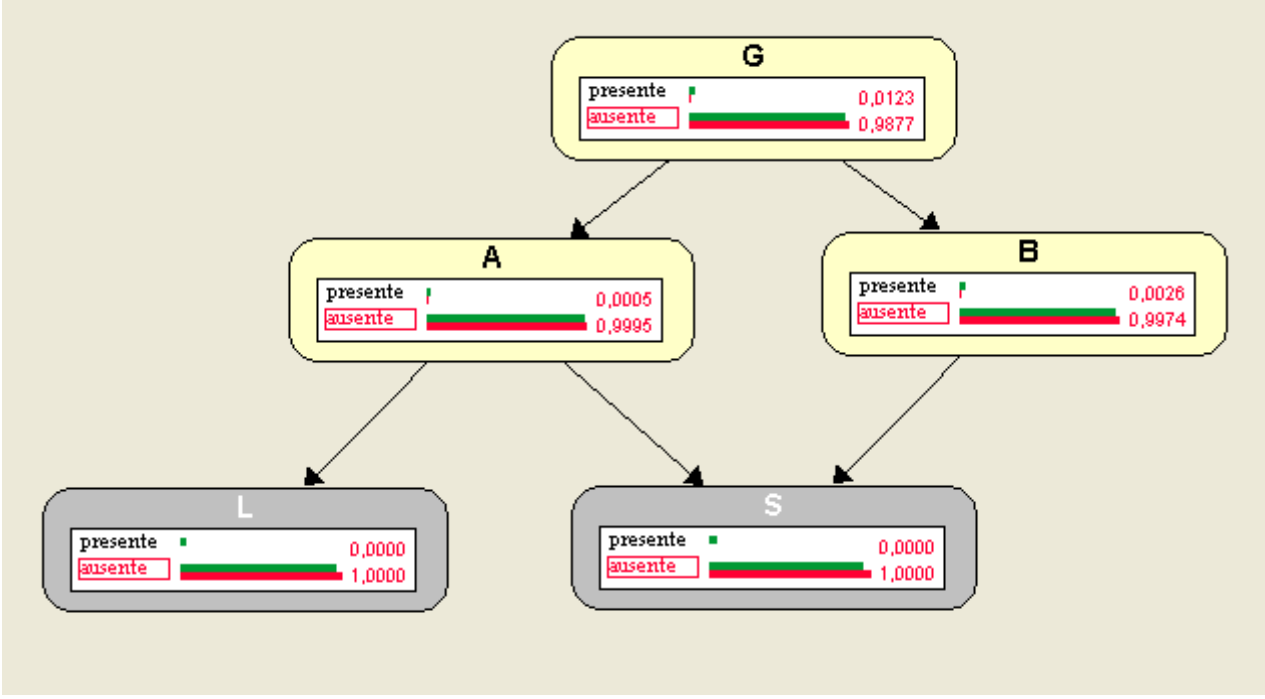



**Imagen en modo inferencia correspondiente al caso** $\boxed{\neg s, \neg l}$



Aunque el método de cálculo es muy distinto, las diferencias numéricas obtenidas entre los resultados de la red bayesiana y los del método probabilístico clásico en este supuesto son muy pequeñas. En el modelo probabilístico clásico, se incluye la información del gen *G* como producto de un test genético convirtiéndolo así en un hallazgo (efecto) de la enfermedad; mientras que en la red bayesiana el gen *G* se ha incluido como causa de las enfermedades.

El siguiente supuesto se modelizará mediante una red bayesiana

**Supuesto 2**. Una mujer decide someterse a un tratamiento de inseminación artificial. Antes de recibir el tratamiento pregunta a su médico por la probabilidad que tiene de quedarse embarazada a lo que él le responde que de un 87%. Unas semana después de la aplicación del tratamiento quiere averiguar si está embarazada, y para ello puede elegir entre tres pruebas de detección del embarazo. La primera prueba es una *ecografía*, que tiene un 1% de probabilidad de ofrecer un falso positivo y un 10% de ofrecer un falso negativo. La segunda es un análisis de sangre, con este test se detectará el nivel de la hormona CGH (gonadotropina coriónica humana) en sangre, y presenta un 10% de probabilidad de ofrecer un falso positivo y un 30% de probabilidad de ofrecer un falso negativo. Y la tercera es un análisis de orina, que, al igual que el análisis de sangre, tratará de detectar el nivel de CGH en la orina, y presenta un 10% de probabilidad de ofrecer un falso positivo y un 20% de probabilidad de ofrecer un falso negativo. La probabilidad de que el nivel de CGH sea detectable en una mujer embarazada es del 90% y la probabilidad de que el nivel de la hormona sea detectable en una mujer no embarazada es del 1% . La paciente quiere hacerse dos pruebas con el objetivo de "buscar mayor certeza" acerca de su estado, pero no sabe que dos pruebas elegir. ¿Qué pareja de tests sería más recomendable para la paciente? Al día siguiente, piensa: "¿Por qué no me hago las tres pruebas?" ¿Qué resultado de las tres pruebas llevaría a la paciente a ser más pesimista en cuanto a su embarazo?

Las variables que intervienen en este problema son:

- Variable ***E*** (Embarazo): $+e\ (sí)$ , $\neg e\ (no)$

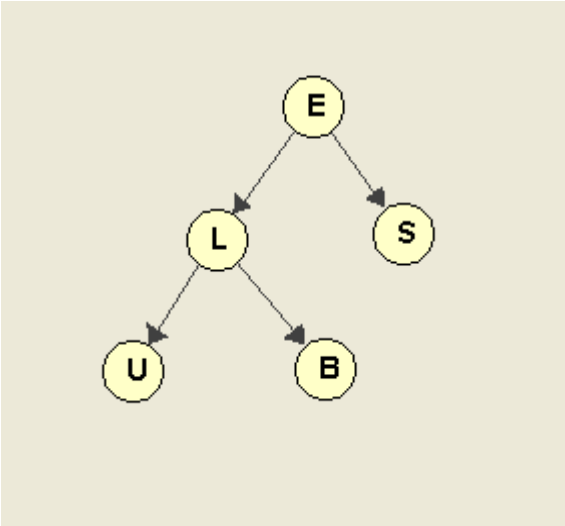
- Variable ***L*** (Nivel GCH): $+l\ (detectable)$ , $\neg l\ (no\,detectable)$

- Variable ***S*** (Ecografía): $+s\ (positivo)$ , $\neg s\ (negativo)$

- Variable ***U*** (Análisis de orina):, $+u\ (positivo)$ , $\neg u\ (negativo)$

- Variable ***B*** (Análisis de sangre): $+b\ (positivo)$ , $\neg b\ (negativo)$

El grafo asociado es:

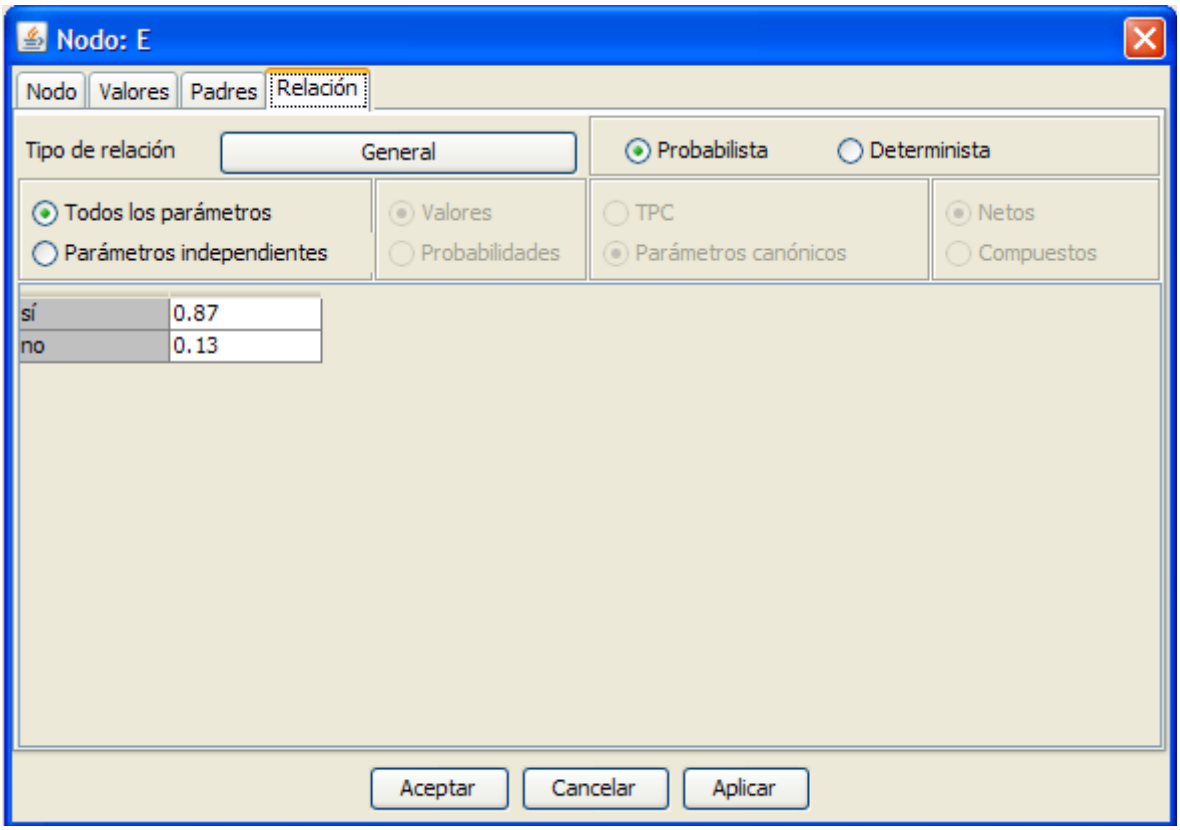



Las relaciones causales establecidas en el grafo se ajustan totalmente a la realidad: la ausencia de enlaces entre los distintos tests (hallazgos), y la independencia existente entre ellos, ya que el resultado de un test no influye en el resultado del otro.

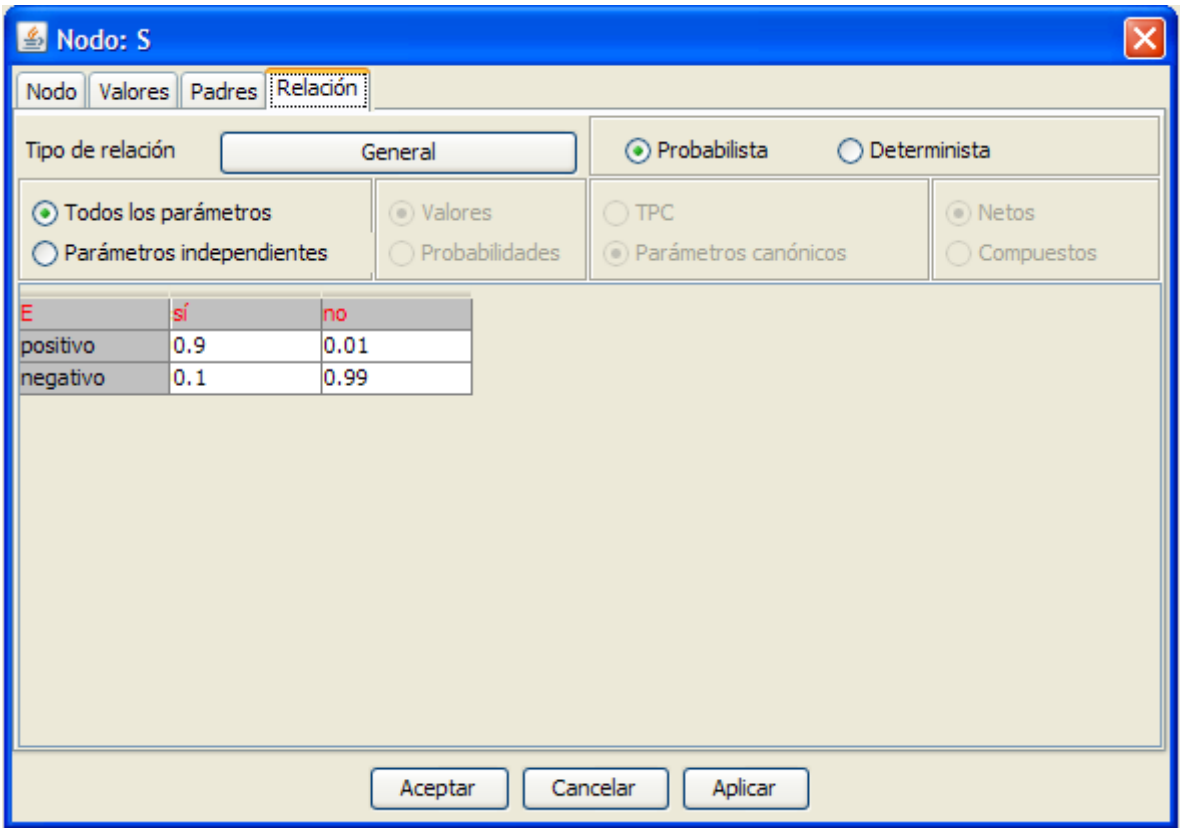
En esta red no se puede aplicar una puerta OR a ningún nodo. Para que esto fuera posible, tendría que tener al menos un nodo que tuviera dos o más padres , pero eso no ocurre ya que el grafo que define la red bayesiana es grafo dirigido acíclico conexo en el que cada nodo tiene como máximo un padre.

Las tablas de probabilidad asociadas a la red bayesiana son las siguientes:

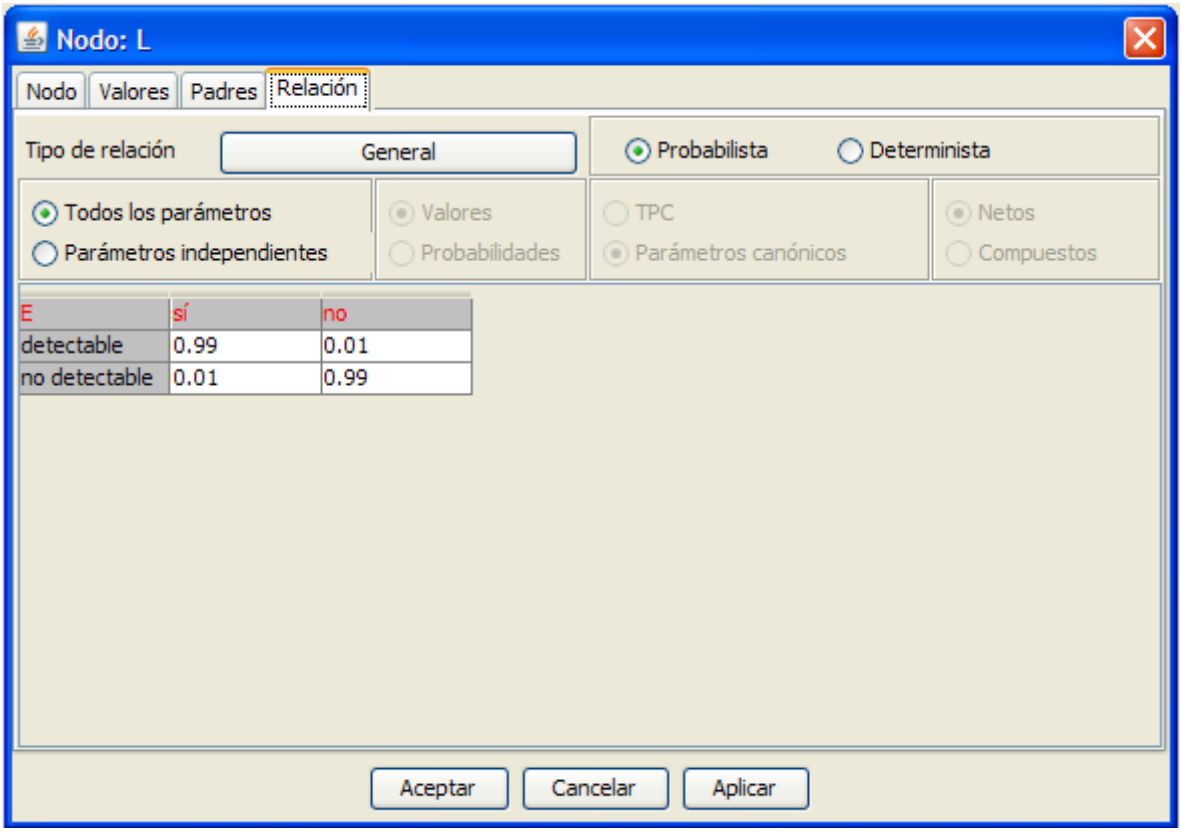
| | $+e$ | $\neg e$ |
|---|---|---|
| $P(E)$ | **0,87** | **0,13** |



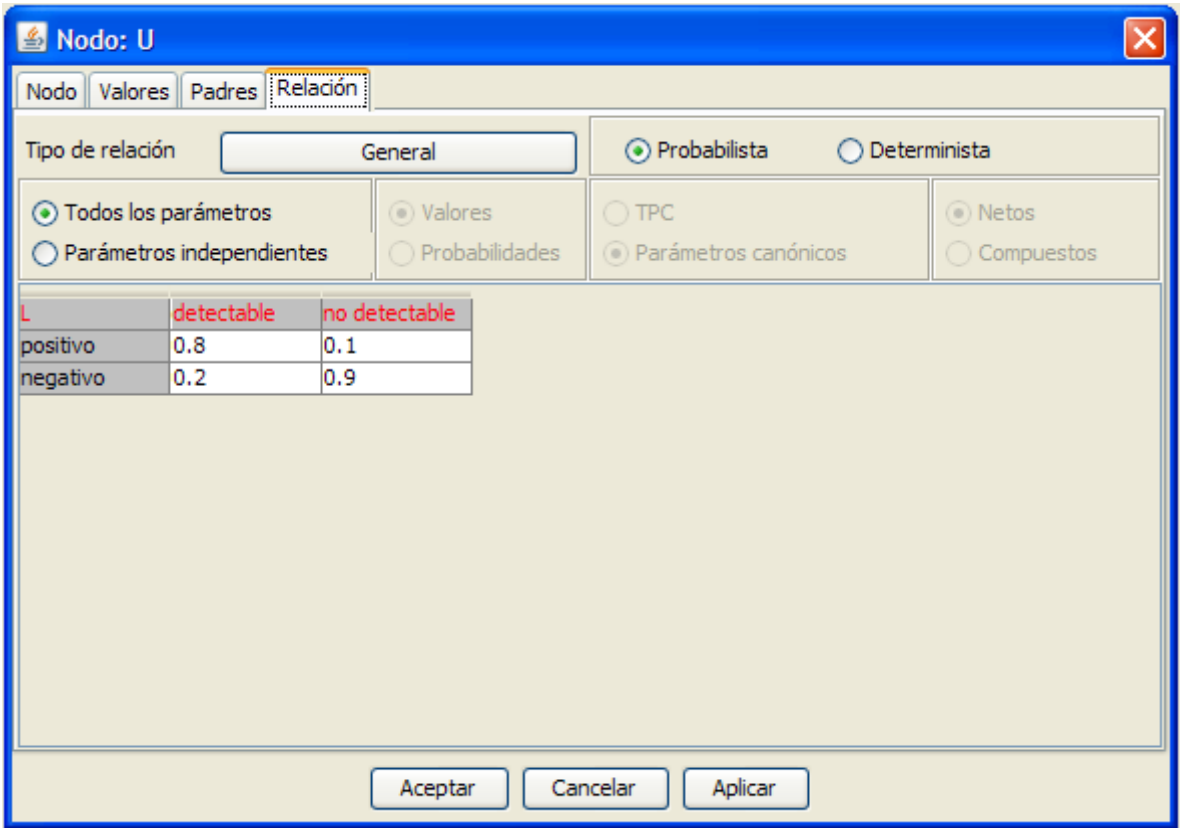

| $P(S\|E)$ | $+e$ | $\neg e$ |
|---|---|---|
| $+s$ | **0,90** | **0,01** |
| $\neg s$ | **0,10** | **0,99** |



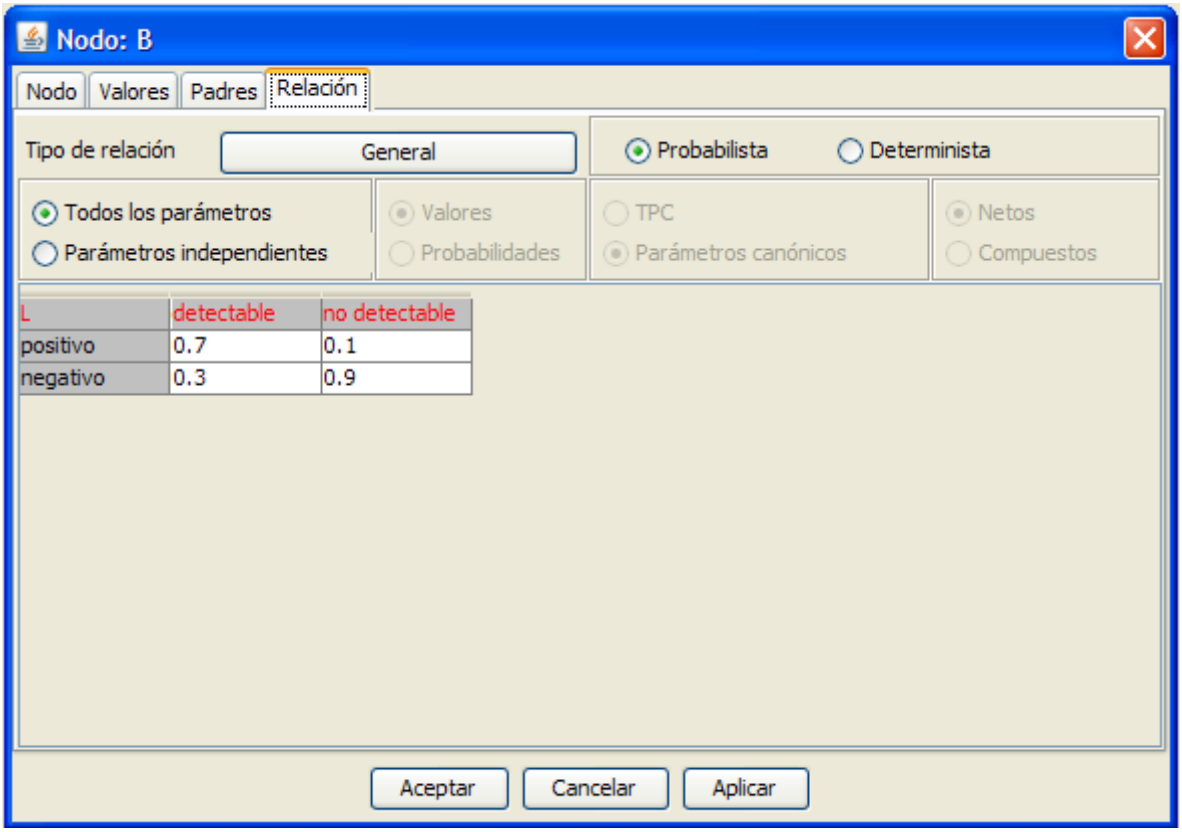

| $P(L\|E)$ | $+e$ | $\neg e$ |
|---|---|---|
| $+l$ | **0,99** | **0,01** |
| $\neg l$ | **0,01** | **0,99** |




| $P(U\|L)$ | $+l$ | $\neg l$ |
|---|---|---|
| $+u$ | **0,80** | **0,10** |
| $\neg u$ | **0,20** | **0,90** |




| $P(B\|L)$ | $+l$ | $\neg l$ |
|---|---|---|
| $+b$ | **0,70** | **0,10** |
| $\neg b$ | **0,30** | **0,90** |




A continuación, se muestra la tabla de probabilidades de $+e$ (estar embarazada) condicionadas a los distintos resultados de los hallazgos. Para elaborar la siguiente tabla se han realizado los cálculos “a mano” de manera análoga a como se hizo en el Supuesto 1, por esa razón no resulta relevante explicitarlos.

| $d$ | $P(+e\|d)$ |
|---|---|
| $+s$ | **1** |
| $\neg s$ | **0,40** |
| $+s, +u$ | **1** |
| $+s, \neg u$ | **0,99** |
| $\neg s, +u$ | **0,83** |
| $\neg s, \neg u$ | **0,14** |
| $+s, +b$ | **1** |
| $+s, \neg b$ | **1** |
| $\neg s, +b$ | **0,82** |
| $\neg s, \neg b$ | **0,19** |
| $+s, +u, +b$ | **1** |
| $+s, \neg u, +b$ | **1** |
| $+s, +u, \neg b$ | **1** |
| $\neg s, +u, +b$ | **0,96** |
| $+s, \neg u, \neg b$ | **0,98** |
| $\neg s, \neg u, +b$ | **0,51** |
| $\neg s, +u, \neg b$ | **0,64** |
| $\neg s, \neg u, \neg b$ | **0,05** |

Se han calculado todos los casos de la tabla anterior con *Elvira*, y como muestra de esta tarea se incluyen las imágenes en modo *Inferencia* correspondientes a algunos casos. Se puede comprobar fácilmente que los cálculos manuales y "automáticos" coinciden.

**Imagen en modo inferencia asociada a** $\neg s, +u$

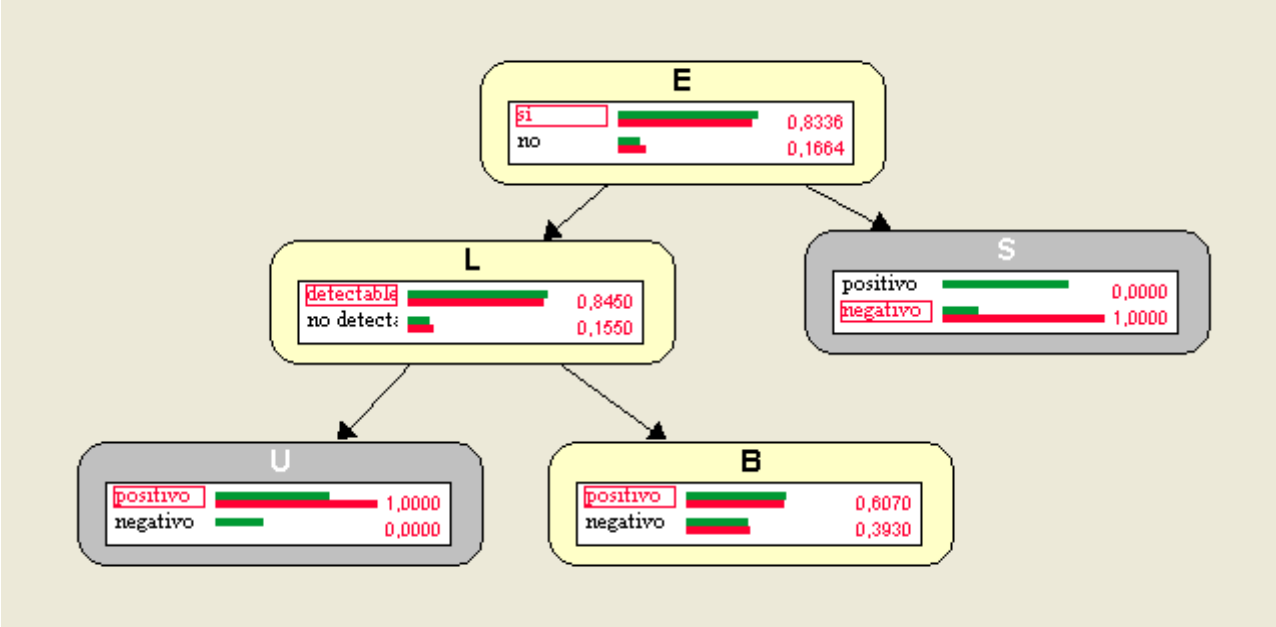



**Imagen en modo inferencia asociada a** $+s, \neg u, +b$

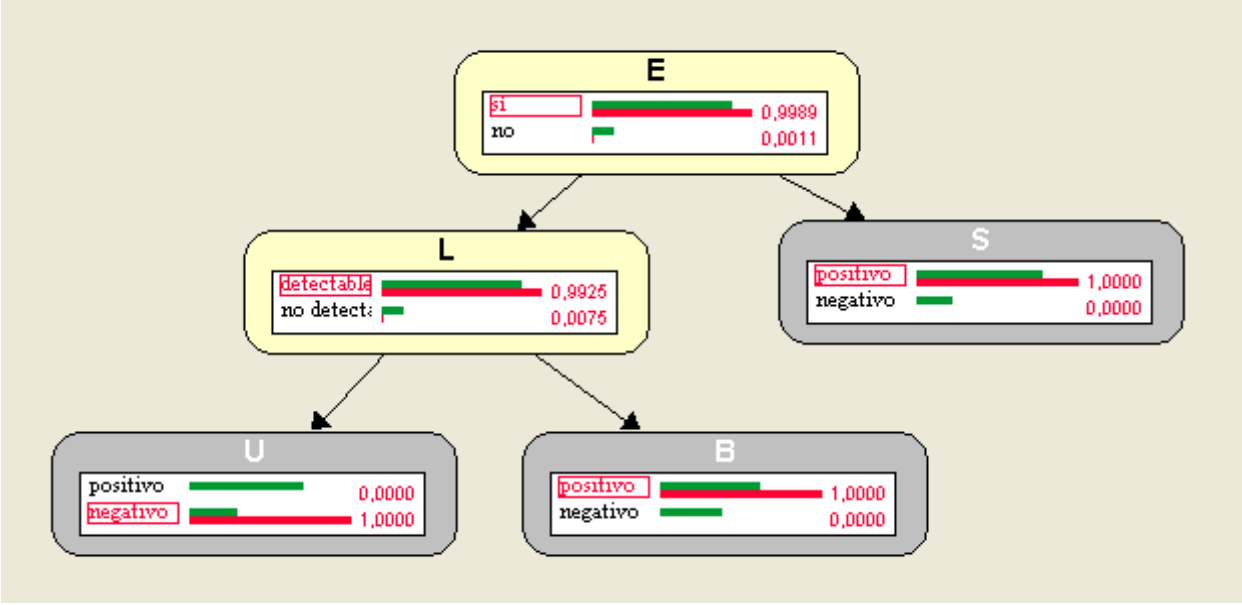


**Imagen en modo inferencia asociada a** $\neg s, +u, +b$

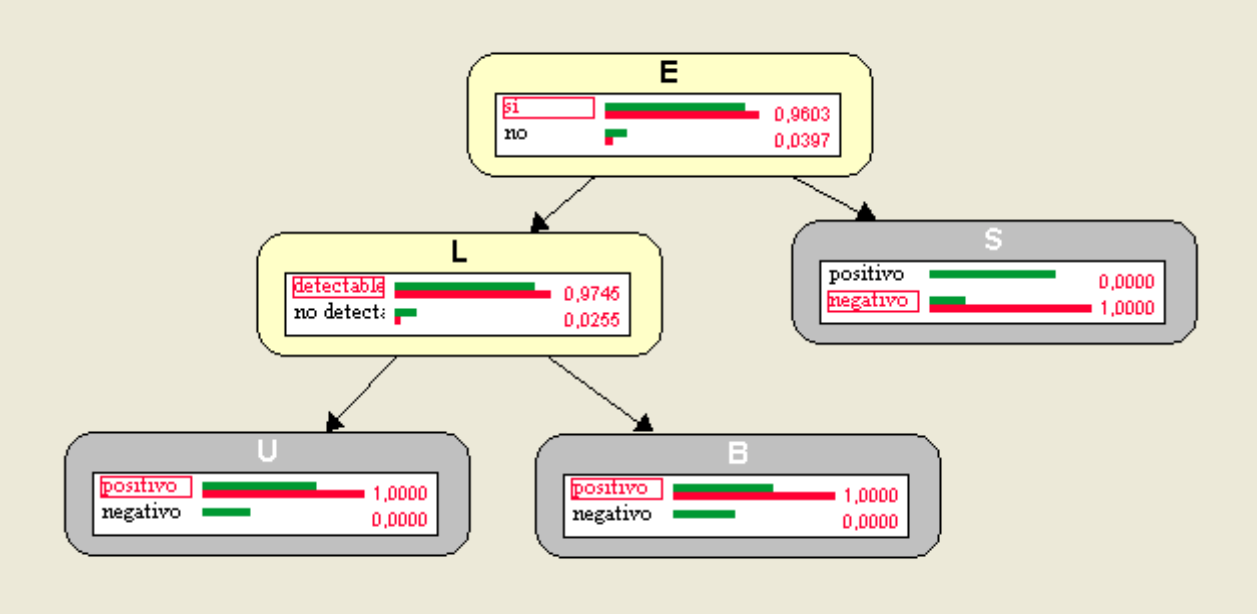



**Imagen en modo inferencia asociada a** $+s, \neg u, \neg b$

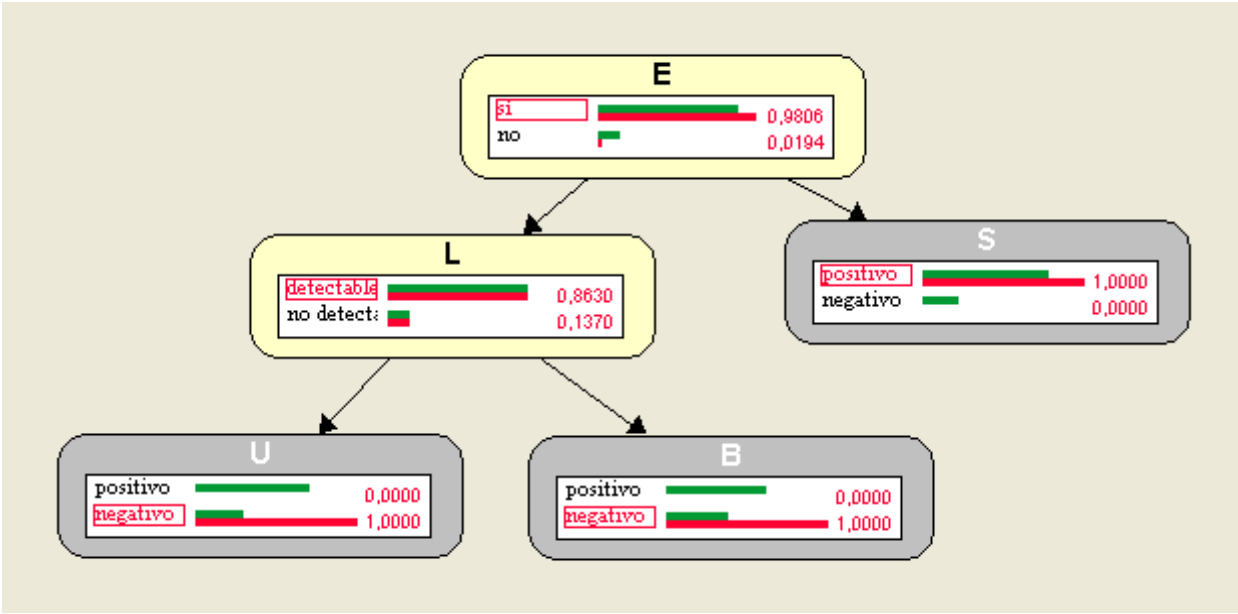
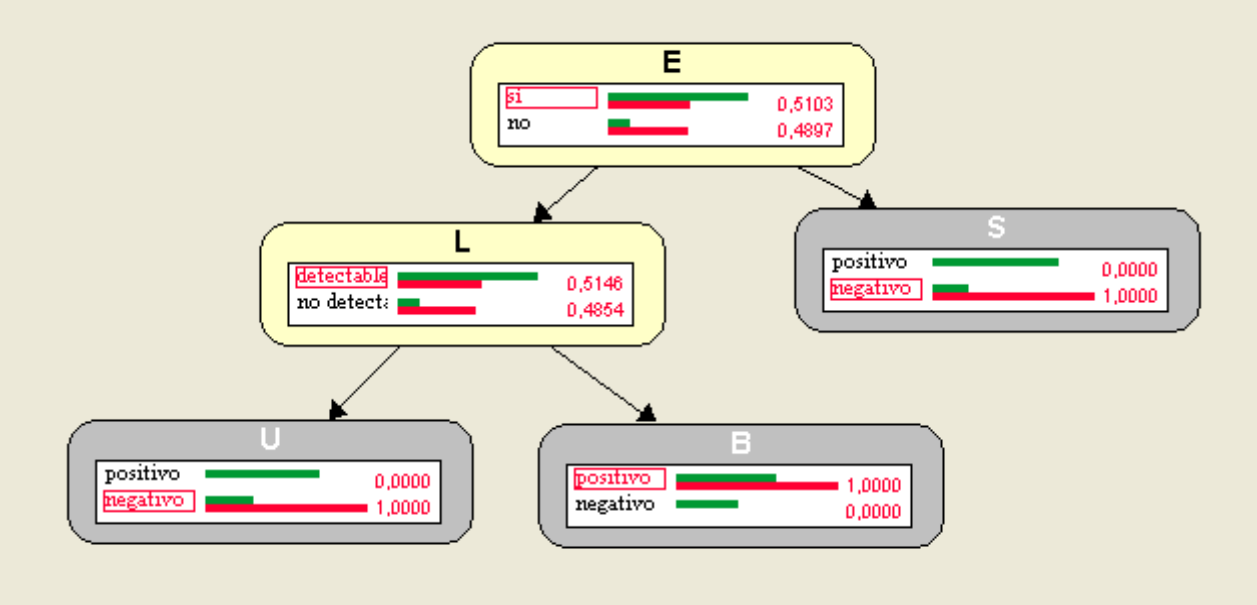


**Imagen en modo inferencia asociada a** $\neg s, \neg u, +b$




# 3. Conclusiones

La red bayesiana constituye una “herramienta” de un valor incalculable en el proceso de toma de decisiones clínicas. La capacidad que posee para representar las relaciones causales que se establecen entre las variables involucradas en un problema, y visualizarlas de una manera muy intuitiva, ha propiciado que se convierta en el modelo gráfico probabilístico más utilizado en el marco de la toma de decisión diagnóstica.

# 4. Bibliografía


1. Bolstad, William M. (2007). *Introduction to Bayesian Statistics*. New Jersey: John Wiley & Sons.

2. Hoff, Peter D. (2007). *A first course in Bayesian Statistical Method*. Heidelberg: Springer.

3. Gelman, A., Carlin, J. B., Stern, Hal S., Dunson, D. B.,Vehtari, A. & Rubin, D. B. (2014). *Bayesian Data Analysis*. Florida: CRC Press.

4. Spiegelhalter, D. J., Abrams, K. R. & Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester: John Wiley & Sons.

5. Jensen, F. V. & Nielsen, T. P. (2007). *Bayesian Networks and Decision Graphs*. New York: Springer.

6. Neapolitan, Richard E. (2004). *Learning Bayesian Networks*. New Jersey: Pearson Prentice Hall.

7. Castillo, E., Gutiérrez, J.M. & Hadi, A.S. (1998). *Sistemas Expertos y Modelos de Redes Probabilísticas*. Madrid: Academia Española de Ingeniería.

8. Machin, D., Campbell, M. J. & Walters, S. J. (2007). *Medical Statistics*. Chichester: John Wiley & Sons.

9. Baron J. (2007). *Thinking and Deciding*. Cambridge: Cambridge University Press.

10. Rius, F. & Wärnberg, J. (2014). *Bioestadística*. Madrid: Paraninfo.

11. Bermejo, Begoña (2001). *Epidemiología clínica aplicada a la toma de decisiones*. Anales del Sistema Sanitario de Navarra (Monografía Nº1). Recuperado de: epidemiología aplicada

12. Scutari, M., Denis, J. B. (2014) *Bayesian Networks: with Examples in R*. Florida: CRC Press.

13. Neapolitan, Richard E. (2009) *Probabilistic Methods for Bioinformatics: with an introduction to Bayesian Networks*. Boston: Morgan Kaufmann.

14. Korb, K. B., Nicholson, A. E. (2011) *Bayesian Artificial Intelligence*. Florida: CRC Press.

15. *Malaria* (2017) [en línea]. World Health Organization. [Consulta: 20 de marzo 2017] <http://www.who.int/malaria/en/>.

16. *Malaria* (2015) [en línea]. MedlinePlus. [Consulta: 20 de marzo 2017] <https://medlineplus.gov/spanish/ency/article/000621.htm>.

17. *Aspectos prácticos del diagnóstico de laboratorio y profilaxis de la Malaria* (2001) [en línea]. Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica: Revisiones Sistemáticas (Parasitología). [Consulta: 20 de marzo 2017] <Diagnóstico Malaria>.

18. *Prueba de embarazo* (2016) [en línea]. MedlinePlus. [Consulta: 10 de abril 2017] <https://medlineplus.gov/spanish/ency/article/003432.htm>.