Xarxes bayesianes: una metodologia per a avaluar riscos

Rosario Delgado
Departament de Matemàtiques
Universitat Autònoma de Barcelona

Resum

En aquest article presentem les xarxes bayesianes, que són models matemàtics probabilístics que es fan servir per modelar situacions en contextos d'incertesa i complexitat. Es poden fer servir, en particular, per a l'avaluació de riscos, és a dir, per estimar la probabilitat de fenòmens amb efectes negatius. Introduïm aquesta metodologia a partir d'exemples i presentem la fórmula de la probabilitat total i la fórmula de Bayes com a eines per actualitzar l'avaluació dels riscos. També parlem de (cor)relacions espúries i factors de confusió, i de com es pot fer servir l'*odds ratio* (o *raó de versemblances*) per valorar el pes de les evidències.

Abstract

In this paper we introduce the Bayesian networks, which are probabilistic mathematical models that are used to model situations in contexts of uncertainty and complexity. They can be used, in particular, for the evaluation of risks, that is to say, to estimate the probability of phenomena with negative effects. We introduce this methodology from examples and present the Total Probability Formula and Bayes' Formula, as tools to update the risk assessment. We also talk about spurious (cor)relationships and confounding factors, and about how the odds ratio (or likelihood ratio) can be used to assess the weight of the evidence.

1. Introducció

En aquest article presentarem les xarxes bayesianes, que són una metodologia que es pot fer servir en contextos on hi ha incertesa per a l'avaluació del risc, és a dir, per estimar la probabilitat d'un fenomen amb efectes negatius. És, per tant, una metodologia probabilística en essència. Consisteix a construir un model matemàtic probabilístic (que acabarà sent estadístic, ja que els paràmetres del model, que són probabilitats, s'estimaran a partir d'una base de dades) per representar les relacions de dependència entre les variables que afecten el fenomen. Un cop construït el model, es pot fer servir per estimar el risc en diferents escenaris i ajudar en la presa de decisions.

Introduirem la metodologia a partir d'exemples i també tractarem altres aspectes relacionats amb aquesta, com són les (cor)relacions espúries i els factors de confusió, que poden representar-se i ser entesos molt millor mitjançant les xarxes bayesianes.

Com acabem de dir, les **xarxes bayesianes** són models matemàtics probabilístics que es fan servir per modelar situacions en contextos d'incertesa i complexitat i, segons l'opinió de molts investigadors del camp de recerca de la **intel·ligència artificial**, la contribució més important en aquest àmbit en els darrers anys ([9]). Trobar la millor manera de raonar en situacions d'incertesa ha estat sempre un tema de preocupació per als científics.

La intel·ligència artificial, si mai arriba a existir en sentit literal, serà una intel·ligència desenvolupada pels humans però implementada com un mecanisme que, sens dubte, haurà de ser capaç de raonar lògicament i de fer front a la **incertesa**. És a dir, haurà de ser capaç de **raonar probabilísticament**, i així podrà tractar amb evidències que porten a creences basades en un coneixement incomplet, a partir de les quals s'arriba a conclusions fal·libles, i podrà reconèixer la necessitat d'aprendre dels errors i de millorar. Per entendre el raonament humà tal com és, és a dir, restringit per la ignorància i la incertesa, l'únic mitjà de què disposem és la teoria de la probabilitat, que ens permet entendre i representar de manera adient la força de les nostres creences i el nostre grau d'incertesa mitjançant la probabilitat, i aplicar les eines de dita teoria per arribar a conclusions.

En una forma molt embrionària, la metodologia de les xarxes bayesianes es va començar a desenvolupar al segle XVII i des de llavors ha estat utilitzada per abordar una gran varietat de problemes que, tanmateix, han hagut de restringir-se quant a la seva mida i complexitat, a causa de la demanda computacional de fer front a problemes de grans dimensions. Fins i tot treballant amb versions simplificades de problemes reals, els humans tenim dificultats per fer els càlculs probabilístics necessaris per resoldrel's. El desenvolupament de les **xarxes bayesianes** facilita aquest procés, automatitzant-lo i evitant així aquestes dificultats. D'altra banda, tant les millores en la capacitat computacional dels ordinadors actuals, com la facultat de les **xarxes bayesianes** d'aprofitar des d'un punt de vista computacional les relacions d'independència entre algunes variables del model, que és un dels seus trets més distintius, permetran en el futur ampliar i aprofundir en l'estudi del raonament probabilístic.

Ja de manera més formal, les xarxes bayesianes van ser introduides en la dècada de 1920, encara que no amb el nom i la nomenclatura actuals, com a eina gràfica que descriu el coneixement probabilístic de les relacions entre variables que afecten un determinat fenomen no determinista. Des de llavors han demostrat la seva enorme utilitat en els procediments de presa de decisions en un gran nombre de camps, i el seu ús per a l'avaluació de riscos està guanyant popularitat dia a dia i inclou aplicacions en àrees tan diverses com l'economia ([1]), la medicina ([18], [4], [20]), el risc mediambiental ([3], [17]), els desastres ecològics ([19]), els accidents amb residus nuclears ([10]) o la criminologia ([5, 6]), per esmentar alguns exemples.

El terme **xarxa bayesiana** i la metodologia en si, ja en el seu format actual, apareixen per primera vegada l'any 1985 en l'article *A Model of Self-Activated Memory for Evidential Reasoning* [15], de Judea Pearl (figura 1 (b)), enginyer, informàtic i filòsof nord-americà, on posa en relleu el paper fonamental que en aquesta metodologia té la **fórmula de Bayes**, introduïda pel matemàtic britànic i ministre presbiterià Thomas Bayes (figura 1 (a)) en el seu famós treball

(a) T. Bayes (Londres, 1702 — Tunbridge Wells, 1761).

(b) J. Pearl (Tel Aviv, 1936).

Figura 1

An Essay towards Solving a Problem in the Doctrine of Chances [2], publicat pòstumament el 1763, justament 222 anys abans que el de Pearl. En aquest treball, Bayes tracta el problema de les causes mitjançant els efectes observats i enuncia la fórmula que porta el seu nom, de la qual parlarem en aquest article, presentant-la com a eina bàsica per tal d'actualitzar les probabilitats, que són una mesura de les nostres creences, a partir de noves evidències.

Juntament amb [15], el text que va marcar una nova era en l'aplicació del raonament probabilístic en la intel·ligència artificial és el llibre del mateix autor, *Probabilistic Reasoning in Intelligent Systems* (1988) [16]. Va influir decisivament en altres investigadors, que es van sentir atrets per aquesta metodologia, com va ser el cas de Richard Neapolitan, que el 1990 va publicar *Probabilistic Reasoning in Expert Systems* [12], un llibre de text que complementa els de Pearl i és una referència bàsica per introduir-se en aquest tema. Aquests textos van establir les xarxes bayesianes com un camp d'estudi perfectament definit i delimitat, amb una gran projecció de futur i una enorme versatilitat i potencialitat quant a les aplicacions als àmbits més variats.

Com ja hem comentat, una de les aplicacions de més èxit de les **xarxes bayesianes** és l'avaluació de riscos. És en aquesta aplicació que ens centrarem en aquest article. A la secció 2 parlarem de què és el risc i com avaluar-lo. Primer presentarem l'aproximació tradicional i després, la nova eina per fer-ho: les xarxes bayesianes. Introduirem tant aquesta metodologia com la **condició de Markov**, la **fórmula de la probabilitat total** i la **fórmula de Bayes**, íntimament relacionades amb ella. Finalment, a la secció 3 parlarem de com es poden presentar les evidències mitjançant l'*odds ratio*, i de la seva interpretació.

L'objectiu de l'article és contribuir que es pugui entendre millor com les probabilitats ens ajuden a representar el nostre coneixement i les relacions entre variables que afecten fenòmens d'interès, i a interpretar la realitat, evitant errors i prejudicis. En definitiva, com ens ajuden a tenir una visió més crítica i a entendre millor el món que ens envolta.

2. Avaluació del risc

2.1. El risc

Segons la definició donada per l'organisme Committee of Sponsoring Organizations of the Treadway Commission[1] el 2004, el **risc** és un esdeveniment que pot tenir conseqüències negatives. En canvi, d'un esdeveniment que pot tenir conseqüències positives en diem una **oportunitat**. Amb un abús del llenguatge, també s'anomena **risc** una mesura numèrica associada a l'esdeveniment que pot tenir conseqüències negatives. És en aquest sentit que farem servir aquest terme a partir d'ara.

Donat un esdeveniment que pot tenir conseqüències negatives, estimar el seu risc associat ens permetrà prendre decisions que ajudin a intentar evitar-ho o a pal·liar les seves conseqüències. Exemples d'esdeveniments pels quals podem estar interessats en estimar el risc serien les catàstrofes naturals o les provocades per l'home. Una catàstrofe és, per definició, un escenari amb una baixa probabilitat de produir-se però amb unes conseqüències desastroses... Segons Craig Fugate, de la Federal Emergency Management Agency dels Estats Units, «treballant ara sobre els plans d'emergència, estarem millor preparats quan necessitem respondre a un esdeveniment així». L'emfàsi és, per tant, en la **prevenció**. I la prevenció passa per una estimació prou acurada del risc que es produeixi la catàstrofe.

Són catàstrofes naturals les erupcions volcàniques, com la del Kilauea, el més actiu dels cinc volcans de l'arxipèlag de Hawaii, que va entrar en erupció el maig del 2018 i, segons els experts, pot continuar durant mesos i, fins i tot, anys, provocant que rius de pedra fosa, fum i cendres corrin pel sud-est de l'illa Gran fins a les aigües del Pacífic. Altres catàstrofes són les tempestes tropicals, els ciclons i els huracans, així com les fortes inundacions o les tempestes de sorra, per esmentar-ne algunes.

A banda d'aquestes catàstrofes que es produeixen al si del nostre planeta, d'altres poden ser provocades des de l'exterior. Per exemple, la Terra es troba sota un bombardeig gairebé constant de partícules carregades que li arriben des de l'espai i que poden ser perjudicials. La principal font d'aquestes partícules és el Sol. Les tempestes (o vents) solars causades per les flamarades solars, que són explosions que es produeixen a la nostra estrella i causen una expulsió de massa de la corona, ens fan arribar radiació electromagnètica que pot arribar massivament a la magnetosfera (zona d'aproximadament 100 Km al voltant de la Terra en la qual el camp magnètic terrestre desvia la major part del vent solar formant un escut protector contra les partícules carregades d'alta energia procedents del Sol) i aconseguir travessar-la, de manera que interferirà greument amb els satèl·lits, els senyals de GPS i les ones de ràdio. El 1859 es va produir una tempesta solar extremament forta en un moment en el qual la humanitat ja estava suficientment desenvolupada per notar les seves conseqüències, encara que no van ser gaire greus: afectacions al telègraf i els aparells elèctrics, i aurores boreals en punts del planeta on no es produïen mai. Aquest fenomen va rebre el nom d'**esdeveniment**

1. Aquest organisme és una iniciativa conjunta per combatre el frau corporatiu que va ser establerta als Estats Units per les cinc organitzacions del sector privat següents: American Accounting Association, American Institute of Certified Public Accountants, Financial Executives International, The Association of Accountants and Financial Professionals in Business i The Institute of Internal Auditors. Es dedica a guiar les administracions executives i les entitats de govern en aspectes rellevants de governança organitzativa, ètica empresarial, control intern, gestió de riscos empresarials, fraus i informes financers.

Carrington en honor del científic anglès que el va descriure. Segons els científics, més tard o més d'hora es tornarà a produir. I ara les seves conseqüències seran molt més greus per a tots nosaltres.

També de l'espai arriben els meteorits. El cràter Manicouagan (figura 2) és possiblement degut a l'impacte d'un meteorit de 5 Km de diàmetre que es va produir fa aproximadament uns 215,5 milions d'anys.

Figura 2. Cràter Manicouagan (Canadà). Fotografia de la NASA.

Hi ha un altre cràter més recent, d'uns 180 Km de diàmetre, a la península mexicana del Yucatán, anomenat Chicxulub, datat fa uns 65 milions d'anys. Molts científics creuen que la caiguda del meteorit que el va crear és la causa de l'extinció dels dinosaures. Es tractaria d'un dels meteorits que s'anomenen «destructors totals».

Podem valorar el risc de patir una catàstrofe? I podem estimar com aquest risc quedaria afectat si es prenen determinades mesures per intentar evitar-la, o per pal·liar les seves conseqüències?

2.2. Aproximació tradicional

La manera tradicional d'estimar el risc és multiplicar la probabilitat de l'esdeveniment amb conseqüències negatives per una mesura del seu impacte (negatiu), segons aquesta fórmula:

$$\text{risc} = \text{probabilitat} \times \text{impacte}$$

Vegem ara amb un exemple senzill que aquesta aproximació és excessivament simplista i ha de ser substituida per una de més elaborada.

Un exemple: *Armaggeddon*

Armaggeddon és el títol d'una pel·lícula del 1998 dirigida per Michael Bay i protagonitzada per Bruce Willis, Ben Affleck i Liv Tyler. També és un terme bíblic que fa referència a la fi catastròfica del món. El seu argument és el següent: els investigadors de la NASA descobreixen un meteorit destrutor total que va directe cap a la Terra i decideixen enviar-hi un grup de perforadors petroliers, dirigit pel personatge interpretat per Bruce Willis (figura 3), per tal de col·locar unes bombes i fer-lo explotar, a fi d'evitar la col·lisió.

Figura 3. L'equip de perforadors de la pel·lícula *Armaggeddon* quan comencen la seva missió.

El món s'enfronta, doncs, a un esdeveniment amb conseqüències molt negatives: la col·lisió del meteorit, una veritable catàstrofe! L'*Armaggeddon* es produiria com a conseqüència de la col·lisió i es podria definir com una pèrdua total al planeta del 80% o més de vides humanes.

Si intetem calcular el risc de col·lisió fent servir l'aproximació tradicional, ens trobem amb algunes dificultats. Per començar, no podem mesurar directament la probabilitat que es produeixi sense aprofundir una mica més... Per exemple, si segons els científics de la NASA la trajectòria del meteorit passa per la Terra, la probabilitat que es produeixi la col·lisió seria 1..., però si és 1, quin sentit té enviar un equip a intentar evitar-la?

De fet, la probabilitat que el meteorit col·lideixi contra la Terra està condicionada per altres esdeveniments (com la intervenció per intentar destruir-lo) i no té sentit assignar-li un valor directament sense tenir-los en compte.

Tampoc no podem obtenir de manera clara una mesura de l'impacte que tindria la col·lisió del meteorit. A banda de la pregunta òbvia: «impacte sobre què?», és clar que no podem mesurar-ho sense considerar les possibles accions mitigants (com ara allotjar la quantitat més gran possible de gent en refugis subterranis tan lluny com sigui possible de la zona d'impacte).

Què podem fer per estimar el risc en aquesta situació? Construirem un model matemàtic probabilístic que permeti incloure tots els esdeveniments que poden condicionar tant la probabilitat que es produeixi la col·lisió del meteorit contra la Terra, com l'impacte que aquest fet tindria.

2.3. Aproximació mitjançant xarxes bayesianes

Si continuem amb l'exemple de l'*Armaggeddon*, ens podem preguntar com serà el model matemàtic probabilístic que farem servir per a avaluar el risc associat a la col·lisió del meteorit contra la Terra. Serà una representació gràfica de les relacions que hi ha entre les diferents variables que són rellevants en aquesta situació, és a dir, variables que afectin el risc associat a la col·lisió del meteorit contra la Terra. Fent-lo servir, podrem realment avaluar aquest risc i prendre mesures.

Primer fem una versió simplificada. Considerem només dues variables per començar: X: Trajectòria de col·lisió (Sí/No), i Y: Col·lisió amb la Terra (Sí/No). La segona variable indica si realment es produeix la col·lisió del meteorit amb el nostre planeta, que és l'esdeveniment de resultat negatiu que ens interessa, i està condicionada per la primera, que es refereix a si la trajectòria actual del meteorit és realment de col·lisió, cas en el qual la col·lisió es produirà, i que actua com a *activador*. En cas contrari, la col·lisió no es produirà. Per tant, tindriem la representació gràfica de la figura 4, amb les probabilitats corresponents.

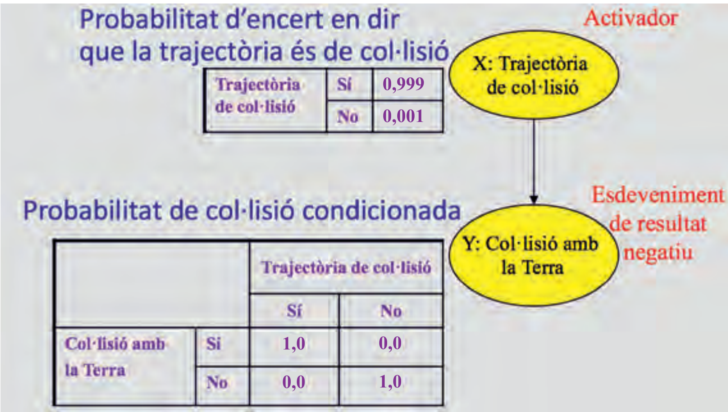
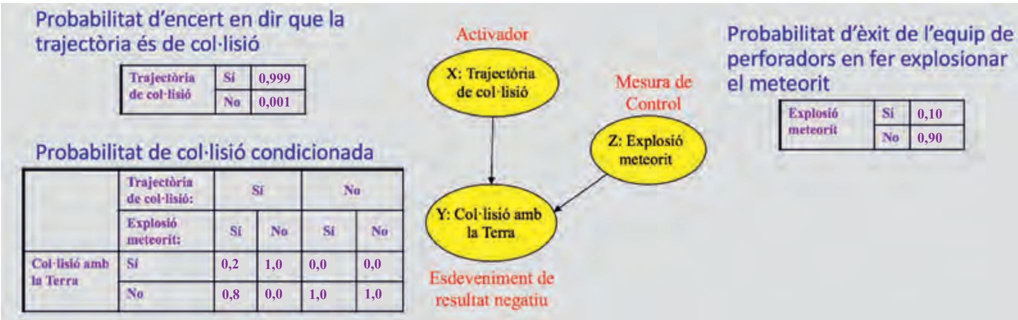


Figura 4. El primer pas en la construcció de la xarxa bayesiana per a l'exemple *Armaggeddon*.

En aquesta representació gràfica hi ha dos nodes que corresponen a les dues variables, amb una fletxa que indica la relació entre elles. Quant a les probabilitats, la taula de probabilitat de la variable Y: Col·lisió amb la Terra condicionada a X: Trajectòria de col·lisió, és prou intuïtiva. La taula de probabilitats de la variable X: Trajectòria de col·lisió, ens indica que creiem que els científics de la NASA han calculat bé la trajectòria del meteorit amb una probabilitat de 999 sobre 1.000, però que amb una probabilitat d'1 sobre 1.000 creiem que s'equivoquen i, en realitat, la trajectòria del meteorit no és de col·lisió contra la Terra.

En un segon pas tindrem en compte també el possible efecte de l'expedició dels perforadors enviats per destruir el meteorit. Introduïm, doncs, una tercera variable: Z: Explosió meteorit (Sí/No), que és una mesura de control i ens indica si l'expedició aconsegueix amb èxit produir l'explosió del meteorit. Aquesta variable afecta la col·lisió del meteorit, però no si la seva trajectòria actual és de col·lisió. Per tant, quan l'afegim al nostre model inicial tindrem el nou model que ens mostra la figura 5, on la probabilitat de la variable Y: Col·lisió amb la Terra ara està condicionada no només per X: Trajectòria de col·lisió, sinó també per Z: Explosió meteorit, que són els seus «pares[2]». En efecte, si la trajectòria no és de col·lisió, passi el que passi amb l'explosió del meteorit, amb probabilitat 1 no es produirà la col·lisió amb la Terra (no considerem la possibilitat d'extrema mala sort consistent que el meteorit inicialment no té una trajectòria de col·lisió amb la Terra però gràcies a l'èxit de l'expedició per fer-lo explotar, es desvia de la seva trajectòria original i la nova trajectòria sí que és de col·lisió contra el nostre planeta). D'altra banda, si la trajectòria és de col·lisió i l'expedició no aconsegueix fer explotar el meteorit, és a dir, fracassa, es produeix la col·lisió amb seguretat (probabilitat 1), mentre que si l'expedició se'n surt, amb una probabilitat del 80% no es produeix la col·lisió amb la Terra, però encara hi ha un 20% de probabilitat que l'efecte no sigui el desitjat i, tot i produir-se l'explosió, el meteorit acabi col·lidint contra nosaltres. A més, creiem que en Bruce Willis i el seu equip tenen només una probabilitat d'èxit del 10% (taula de probabilitat de la variable Z en la figura 5). Sembla poc, però les condicions de la tasca són molt extremes, com es pot comprovar si es veu la pel·lícula!



Trajectòria de col·lisió	Sí	0,999
	No	0,001

	Trajectòria de col·lisió:	Sí		No	
	Explosió meteorit:	Sí	No	Sí	No
Col·lisió amb la Terra	Sí	0,2	1,0	0,0	0,0
	No	0,8	0,0	1,0	1,0

Explosió meteorit	Sí	0,10
	No	0,90

Figura 5. Segon pas en la construcció de la xarxa bayesiana per a l'exemple *Armaggeddon*.

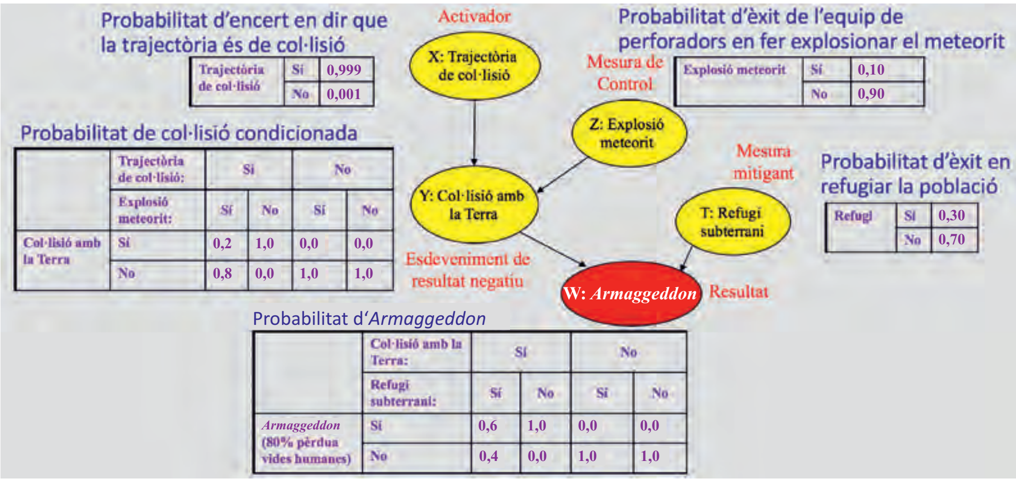
Finalment, acabem de contruir el model considerant l'efecte mitigant de les conseqüències negatives de la col·lisió del meteorit que s'obtindria refugiant la població sota terra, amb la variable T: Refugi subterrani (Sí/No), que indica si la mesura mitigant de refugi de la població s'ha pogut implementar amb èxit. El resultat negatiu de l'esdeveniment de la col·lisió del meteorit amb el planeta seria la pèrdua de vides humanes, que si és massiva (almenys un 80% de la població mundial) anomenem *Armaggeddon*. Aquest resultat es representa per la variable W: Armaggeddon (Sí/No). A la figura 6 veiem el model final, amb totes les variables incorporades.

Només hem afegit les dues noves variables T i W, tenint en compte que T: Refugi subterrani només afecta W: *Armaggeddon* (i no la resta de variables), de la qual és pare juntament amb Y: Col·lisió amb la Terra. Per tant, la taula de probabilitat condicionada de la variable W: *Armag-*

2. En la terminologia habitual de les xarxes bayesianes, en diem *pares* d'una variable o node als nodes que li envien una fletxa en la representació gràfica, dels quals depèn directament i n'és un *fill*. Quan un node no té pares, en diem que és un node *arrel*; quan no té *fills*, és a dir, nodes dels quals sigui pare, en diem que és un node *fulla*; i de la resta de nodes en diem que són *intermedis*. Del conjunt de nodes als quals es pot anar seguint en la direcció de les fletxes des d'un node fixat, se'n diu *descendents* del node, mentre que els seus *ancestres* són els nodes als quals es pot anar seguint en el sentit invers de les fletxes, és a dir, els nodes dels quals és *descendent*.

geddon depèn de Y: Col·lisió amb la Terra i de T: Refugi subterrani, i de la resta de variables del model només a través de la primera. Si no es produeix la col·lisió, independentment de l'èxit de la mesura mitigant, no es produirà l'*Armaggeddon* amb probabilitat 1. En canvi, si es produeix la col·lisió, la probabilitat d'*Armaggeddon* dependrà de l'èxit de la mesura mitigant: si fracassa, l'*Armaggeddon* és segur (probabilitat 1), mentre que si té èxit, només es produeix amb probabilitat del 60%. D'altra banda, creiem que la probabilitat que la mesura del refugi subterrani es pugui acabar implementant és del 30% (taula de probabilitat de la variable T en la figura 6).

El model que acabem de construir (figura 6), és una **xarxa bayesiana**.



Trajectòria de col·lisió		
	Sí	0,999
	No	0,001

Explosió meteorit		
	Sí	0,10
	No	0,90

	Trajectòria de col·lisió:	Sí		No	
	Explosió meteorit:	Sí	No	Sí	No
Col·lisió amb la Terra	Sí	0,2	1,0	0,0	0,0
	No	0,8	0,0	1,0	1,0

Refugi		
	Sí	0,30
	No	0,70

	Col·lisió amb la Terra:	Sí		No	
	Refugi subterrani:	Sí	No	Sí	No
Armaggeddon (80% pèrdua vides humanes)	Sí	0,6	1,0	0,0	0,0
	No	0,4	0,0	1,0	1,0

Figura 6. Pas final en la construcció de la xarxa bayesiana per a l'exemple *Armaggeddon*.

2.4. Les xarxes bayesianes i la condició de Markov

Com hem vist en l'exemple de l'*Armaggeddon*, una xarxa bayesiana és un model matemàtic probabilístic que representa les relacions (sotmeses a l'atzar) entre unes variables, diguem $X_1, \ldots, X_n$, rellevants per a cert esdeveniment (de conseqüències negatives, en el cas de l'avaluació de riscos) i que són variables **discretes finites** (numèriques que prenen un conjunt finit de possibles valors) o **categòriques** (els valors de les quals són un conjunt finit de categories). El model consta d'una representació gràfica que és un graf acíclic dirigit (DAG, les sigles en anglès de *Directed Acyclic Graph*), que denotem per Γ, ja que les relacions entre les variables s'estableixen amb arcs que tenen una direcció (fletxes que van d'un node a un altre) i no està permès que, sortint d'un node i seguint el sentit de les fletxes, anem passant de node a node i tornem al de partida (si això passés, es tindria un cicle tancat). A més de la representació gràfica, tenim els paràmetres del model, que són les probabilitats condicionades de cada node als seus pares (incloent-hi les probabilitats sense condicionar dels nodes arrels). A partir d'aquestes probabilitats, podem definir una distribució de probabilitat conjunta de les variables del model de la manera següent (se'n diu **regla de la cadena**): per a tot $x_1, \ldots, x_n$ possibles valors de $X_1, \ldots, X_n$, respectivament, tenim:

$$\boxed{P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i/PA(X_i)),} \tag{1}$$

on $P(X_i = x_i/PA(X_i))$ denota la probabilitat que la variable X_i prengui el valor x_i condicionada que els seus pares prenguin els valors que correspongui dels $x_1, \ldots, x_n$. Dit d'una altra manera: la probabilitat conjunta de les variables del model es calcula multiplicant les probabilitats condicionades de cada node als seus pares, començant pels nodes *fulla* i anant enrere en el DAG fins a acabar als nodes *arrel*.

Vegem-ho en l'exemple de l'*Armaggeddon*: suposem que volem calcular la probabilitat conjunta de totes les variables del model que correspon al fet que totes prenguin el valor Sí (anàlogament es faria amb qualsevol combinació de Sí i No per a les diferents variables). Aplicant la **regla de la cadena**, aquesta probabilitat es calcula així:

$$\begin{aligned}
&P(\mathrm{X = Sí, Y = Sí, Z = Sí, T = Sí, W = Sí})\\
&= P(\mathrm{W = Sí/Y = Sí,\ T = Sí})\, P(\mathrm{T = Sí})\, P(\mathrm{Y = Sí/X = Sí, Z = Sí})\, P(\mathrm{X = Sí})\, P(\mathrm{Z = Sí})\\
&= 0{,}6 \times 0{,}3 \times 0{,}2 \times 0{,}999 \times 0{,}1 = 0{,}0035964
\end{aligned}$$

Donat el DAG, la distribució de probabilitat conjunta definida d'aquesta manera verifica una propietat que s'anomena **condició de Markov**. Habitualment, aquest terme es refereix a la propietat de certs processos estocàstics (o aleatoris)[3] que evolucionen en el temps i que «no tenen memòria». És a dir, processos pels quals conegut el present, el futur és independent del passat. Aquests processos reben el nom de **cadenes de Markov** en certs casos particulars i, en general, el de **processos de Markov** ([7]), en honor del gran matemàtic rus Andréi Andréyevich Márkov (figura 7).

Figura 7. A.A. Márkov (Ryazan, 1856 — Sant Petersburg, 1922), matemàtic rus conegut pels seus treballs en la teoria de la probabilitat.

3. La paraula estocàstic prové del grec *stockhastikos*, que és una paraula derivada de *stockhos*, que fa referència a l'atzar. La paraula aleatori prové del llatí *aleatorius*, paraula derivada de *alea*, que es refereix al joc dels daus i, en general, a l'atzar. Ambdues paraules es poden fer servir, indistintament, per referir-se a l'atzar. El costum ha fet que en determinats contextos s'utilitzi l'una o l'altra.

En el context dels DAG, la **condició de Markov** s'expressa d'aquesta altra manera: cada variable és condicionalment independent dels nodes que no són descendents seus, atès que es coneix l'estat dels seus pares.

Aquesta condició implica que en l'exemple de l'*Armaggeddon*, la variable W: *Armaggeddon* és independent de les variables X: Trajectòria de col·lisió i Z: Explosió meteorit si coneixem els valors de les variables Y: Col·lisió amb la Terra i T: Refugi subterrani, per posar un exemple.

Tenim el resultat següent:

Teorema (*Theorem 1.5 [13]*)
Si Γ és un DAG sobre el conjunt de variables $V = \{X_1, \ldots, X_n\}$ i P és la distribució de probabilitat conjunta sobre V definida per (1), llavors el parell (Γ, P) verifica la **condició de Markov**.

(Cor)relacions espúries i factors de confusió

Les xarxes bayesianes ens permeten modelar i entendre un fenomen curiós que apareix quan dues variables que, atenent el sentit comú no tindrien cap relació, se'ns mostren com a (cor)relaciones entre elles. Es tracta del que s'anomena (cor)relació espúria. Això succeeix perquè hi ha una tercera variable (que s'anomena factor de confusió) que afecta ambdues però que no és evident i ens confon (per això s'anomena d'aquesta manera). Gràcies a les xarxes bayesianes entendrem aquest fenomen i evitarem caure en paranys. Posem-ne un exemple:

Exemple: la venda de paraigües incrementa els accidents de trànsit. Observem que en un determinat indret els dies que es venen més paraigües també és més probable que hi hagi accidents de trànsit. Clarament, les variables F: Venda paraigües (Alta/Baixa) i G: Accident trànsit (Sí/No) no sembla que tinguin gaire a veure. Com s'explica, doncs, que estiguin (cor)relaciones? S'explica per un factor extern, que és la pluja, causa comuna de les dues variables. En efecte, la pluja afecta tant la venda de paraigües com la sinestralitat de trànsit, i les fa augmentar les dues. Podem modelitzar aquesta situació mitjançant la xarxa bayesiana de la figura 8, en la qual representem la causa comuna C: Pluja (Sí/No) i els seus dos efectes.

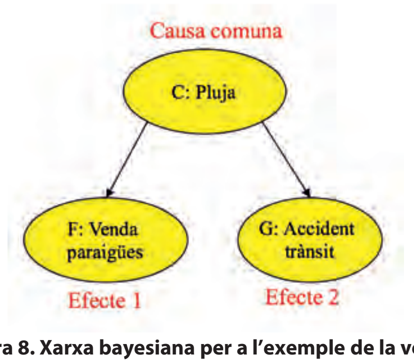


Figura 8. Xarxa bayesiana per a l'exemple de la venda de paraigües que incrementa els accidents de trànsit.

Són independents els dos efectes? No ho són, tot i que cap d'ells és causa de l'altre. Però, si un no és causa de l'altre, com s'influeixen entre ells? Mitjançant la causa comuna. En efecte, si un dia la venda de paraigües és alta, serà més probable que sigui un dia plujós, i si ho és, a causa de la pluja augmentarà la probabilitat que hi hagi accidents de trànsit. Per tant, un determinat valor d'un dels efectes (Venda paraigües = Alta) afecta la probabilitat amb què l'altre efecte pren els seus valors. Podem dir que hi ha una transmissió d'informació d'un dels efectes a l'altre a través de la causa comuna: de l'efecte 1 es transmet la informació a la causa, i de la causa, a l'efecte 2. Ara bé, si fixem el valor de la causa, per exemple, considerem un dia en què sabem que plou, la transmissió d'informació entre els efectes queda tallada i es fan independents, ja que, encara que sabéssim que Venda paraigües = Alta, això no pot fer augmentar la probabilitat de pluja (ja sabem que està plovent!) i, per tant, no es modificarà la probabilitat que es produeixin accidents de trànsit. Fixem-nos que això no és més que la **condició de Markov** aplicada a aquest cas: com que coneixem l'estat del pare (variable C: Pluja), els dos fills són independents entre si ja que cap d'ells no és descendent de l'altre.

Una situació diferent però també interessant és la de dues causes independents però amb un efecte comú, que considerem en l'exemple següent.

Exemple: la gespa del jardí està mullada. Tenim un jardí amb gespa i reg automàtic que es posa en funcionament cada tres dies, faci el temps que faci. Per tant, les variables C: Pluja (Sí/No) i D: Reg automàtic (Sí/No), que fan referència a si un determinat dia plou i a si el reg automàtic s'ha posat en funcionament, són clarament independents. I, tanmateix, podríem crear una dependència entre elles mitjançant una tercera variable? Sí, sempre que sigui un efecte comú, com ara que la gespa del jardí estigui mullada. Fem servir la xarxa bayesiana de la figura 9 per representar la situació, on, a més de les dues variables inicials, tenim la variable efecte comú F: Gespa mullada (Sí/No).

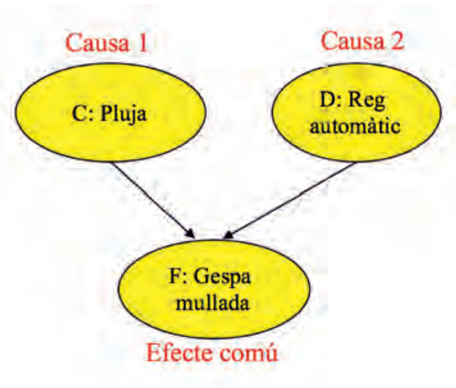


Figura 9. Xarxa bayesiana per a l'exemple de la gespa mullada.

Fixem l'efecte comú, és a dir, fixem un dia en el qual ens adonem que la gespa està mullada. Si és un dia en què no plou, com que una de les causes de l'efecte que hem observat no es dona, això implica que la probabilitat que es doni l'altra causa augmenta, és a dir, augmenta la probabilitat que el reg automàtic s'hagi posat en funcionament. De fet, si estem segurs que no hi ha cap altra causa possible de l'efecte que la gespa estigui mullada, l'absència d'una de les causes obliga a la presència de l'altra. En aquest cas, l'absència de pluja implica necessàriament que el reg automàtic s'ha posat en funcionament i, per tant, si coneixem que F = Sí, llavors C i D **no** són independents.

2.5. Avaluant el risc *a priori*: la fórmula de la probabilitat total

El risc *a priori* és la probabilitat que es produeixin les conseqüències negatives de determinat esdeveniment, si no es disposa de cap més informació addicional. Reprenent l'exemple de l'*Armageddon*, el risc *a priori* és la probabilitat $P(Armaggeddon = Sí)$. Per calcular aquesta probabilitat a partir dels paràmetres del model hem de tenir en compte que depèn dels valors que prenguin les seves variables pares: Y: Col·lisió amb la Terra i T: Refugi subterrani.

Ens ajudem de la figura 10 per calcular $P(Armaggeddon = Sí)$. Aquesta probabilitat es pot representar (és només una representació!) com l'àrea del conjunt A dins un rectangle d'àrea igual a 1, que representa la totalitat. Per tant, la probabilitat és la suma de quatre àrees, les de A_1, A_2, A_3 i A_4, cadascuna d'elles corresponent a la intersecció entre A i una de les quatre parts en què s'ha dividit el rectangle d'àrea 1, B_1, B_2, B_3 i B_4. Aquestes quatre parts venen donades per les combinacions de possibles valors dels dos pares. És a dir:

$$P(Armaggeddon = Sí) = P(A) = P(A_1) + P(A_2) + P(A_3) + P(A_4),$$

i cadascuna d'aquestes probabilitats es calcula de la manera següent:

$$\boxed{P(A_i) = P(A \cap B_i) = P(A/B_i)\,P(B_i).} \tag{2}$$

Aquesta darrera fórmula per calcular la probabilitat de la intersecció de dos esdeveniments a partir de la probabilitat d'un d'ells condicionada a l'altre es pot justificar així: $P(A_i) = P(A \cap B_i)$ és la proporció d'àrea que representa la porció A_i respecte del rectangle total i coincideix amb el producte de $P(A/B_i)$, que és la proporció d'àrea que representa A respecte de B_i, i de $P(B_i)$, que és la proporció que representa B_i respecte del total. Per exemple, si l'àrea de B_i és $1/4$ del total i l'àrea d'A representa $1/3$ de l'àrea de B_i, llavors $A_i = A \cap B_i$ té una àrea que representa $1/3 \times 1/4 = 1/12$ del total.

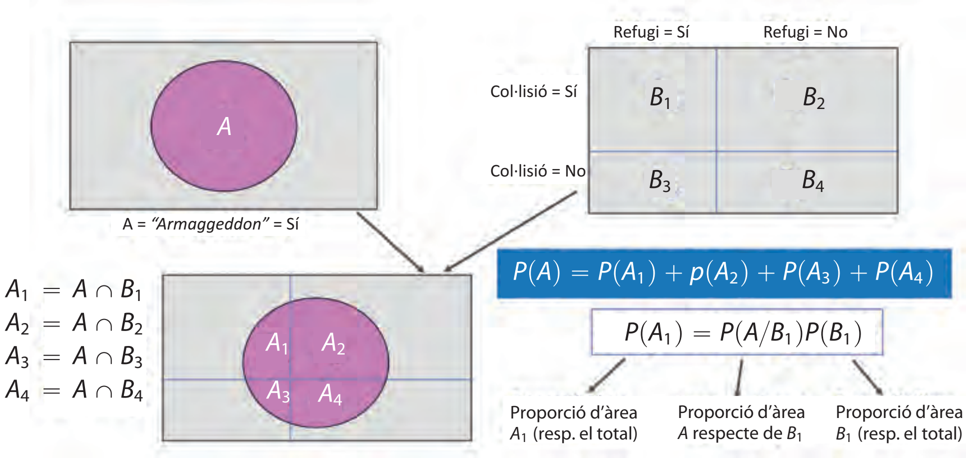


Figura 10. Cap a la fórmula de la probabilitat total.

En general, tenim que si dividim el rectangle que representa el total en n parts, $B_1, \ldots, B_n$, la probabilitat de l'esdeveniment A es pot calcular sumant les probabilitats d'intersecció d'A amb cadascuna de les parts; és a dir:

$$\boxed{P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A/B_i)\, P(B_i).} \tag{3}$$

Aquesta expressió es coneix com a **fórmula de la probabilitat total** i, aplicant-la en el nostre cas, podem calcular el risc *a priori* de la manera següent (fent servir les probabilitats de la figura 6):

$$\begin{aligned} P(\mathrm{W} = \mathrm{Sí}) = {} & P(\mathrm{W} = \mathrm{Sí}/\mathrm{Y} = \mathrm{Sí}, \mathrm{T} = \mathrm{Sí}) P(\mathrm{Y} = \mathrm{Sí}, \mathrm{T} = \mathrm{Sí}) \\ & + P(\mathrm{W} = \mathrm{Sí}/\mathrm{Y} = \mathrm{Sí}, \mathrm{T} = \mathrm{No})\, P(\mathrm{Y} = \mathrm{Sí}, \mathrm{T} = \mathrm{No}) \\ & + P(\mathrm{W} = \mathrm{Sí}/\mathrm{Y} = \mathrm{No}, \mathrm{T} = \mathrm{Sí})\, P(\mathrm{Y} = \mathrm{No}, \mathrm{T} = \mathrm{Sí}) \\ & + P(\mathrm{W} = \mathrm{Sí}/\mathrm{Y} = \mathrm{No}, \mathrm{T} = \mathrm{No})\, P(\mathrm{Y} = \mathrm{No}, \mathrm{T} = \mathrm{No}) \\ & = 0{,}6 \times P(\mathrm{Y} = \mathrm{Sí}) \times 0{,}3 + 1{,}0 \times P(\mathrm{Y} = \mathrm{Sí}) \times 0{,}7 \\ & + 0{,}0 \times P(\mathrm{Y} = \mathrm{No}) \times 0{,}3 + 0{,}0 \times P(\mathrm{Y} = \mathrm{No}) \times 0{,}7 = 0{,}88 \times P(\mathrm{Y} = \mathrm{Sí}). \end{aligned} \tag{4}$$

Observeu que les probabilitats conjuntes de les variables Y: Col·lisió amb la Terra i T: Refugi subterrani, per exemple $P(\mathrm{Y} = \mathrm{Sí}, \mathrm{T} = \mathrm{Sí})$ (i el mateix amb les altres tres probabilitats), es calculen fent servir la condició de Markov, que ens assegura que són variables independents. En efecte, condicionada als seus pares (que no en té o bé, el que és el mateix en aquest cas, sense condicionar), la variable T: Refugi subterrani és independent de Y: Col·lisió amb al Terra, ja que aquesta no és un descendent seu. Així, com que les dues variables són independents, tenim que la probabilitat conjunta de les dues variables és producte de probabilitats. Això es pot justificar fàcilment a partir de la fórmula (2): per definició, dues variables discretes U i V són independents si (i només si) qualsevol parella d'esdeveniments de la forma $\{U = u\}$ i $\{V = v\}$ són independents i, en aquest cas, per la fórmula (2),

$$P(U = u, V = v) = P(U = u/V = v) = P(V = v) = P(U = u)\, P(V = v)\,,$$

on en la darrera igualtat hem fet servir la independència dels esdeveniments $\{U = u\}$ i $\{V = v\}$ (la qual cosa implica que condicionar per un d'ells no afecta la probabilitat de l'altre). Així, doncs:

$$P(\mathrm{Y} = \mathrm{Sí}, \mathrm{T} = \mathrm{Sí}) = P(\mathrm{Y} = \mathrm{Sí}) = P(\mathrm{T} = \mathrm{Sí}) = P(\mathrm{Y} = \mathrm{Sí}) \times 0{,}3\,.$$

Per obtenir el valor numèric final a (4) necessitem calcular $P(\mathrm{Y} = \mathrm{Sí})$, i això ho fem, de nou, amb la fórmula de la probabilitat total, condicionant els possibles valors dels pares de la variable Y: Col·lisió amb la Terra, que són X: Trajectòria de col·lisió i Z: Explosió meteorit. Així, aplicant de nou la condició de Markov, que ens diu que les variables pares són independents i, per tant, la probabilitat conjunta és producte de probabilitats, tenim que

$$\begin{aligned}
P(\mathrm{Y} = \mathrm{Sí}) &= P(\mathrm{Y} = \mathrm{Sí}/\mathrm{X} = \mathrm{Sí}, \mathrm{Z} = \mathrm{Sí})\, P(\mathrm{X} = \mathrm{Sí}, \mathrm{Z} = \mathrm{Sí}) \\
&+ P(\mathrm{Y} = \mathrm{Sí}/\mathrm{X} = \mathrm{Sí}, \mathrm{Z} = \mathrm{No})\, P(\mathrm{X} = \mathrm{Sí}, \mathrm{Z} = \mathrm{No}) \\
&+ P(\mathrm{Y} = \mathrm{Sí}/\mathrm{X} = \mathrm{No}, \mathrm{Z} = \mathrm{Sí})\, P(\mathrm{X} = \mathrm{No}, \mathrm{Z} = \mathrm{Sí}) \\
&+ P(\mathrm{Y} = \mathrm{Sí}/\mathrm{X} = \mathrm{No}, \mathrm{Z} = \mathrm{No})\, P(\mathrm{X} = \mathrm{No}, \mathrm{Z} = \mathrm{No}) \\
&= 0{,}2 \times 0{,}999 \times 0{,}1 + 1{,}0 \times 0{,}999 \times 0{,}9 \\
&+ 0{,}0 \times 0{,}001 \times 0{,}1 + 0{,}0 \times 0{,}001 \times 0{,}9 = 0{,}91908\,.
\end{aligned}$$

Per tant, finalment, el **risc** *a priori* s'obté susbstituint aquest valor a (4):

$$P(\mathit{Armaggeddon} = \mathrm{Sí}) = 0{,}88 \times 0{,}91908 = 0{,}8087904 \approx 81\%$$

Observació: hem calculat aquesta probabilitat fent servir la fórmula de la probabilitat total, juntament amb la condició de Markov. Exactament de la mateixa manera que si l'haguéssim calculat fent servir la regla de la cadena (1).

2.6. Avaluant el risc *a posteriori*: la fórmula de la probabilitat total condicionada

Si ara disposem d'una evidència, és a dir, d'una informació rellevant per a l'avaluació del risc, podem actualitzar el seu valor obtenint un risc *a posteriori*. Per exemple, imaginem que l'evidència que tenim és que l'equip de perforadors ha fracassat en el seu intent de fer explotar el meteorit. Resulta obvi que aquesta informació modificarà el risc a l'alça, però quant augmenta el risc el fet de conèixer aquesta informació? Per contestar a la pregunta hem de calcular la probabilitat que la variable *Armaggeddon* sigui Sí condicionada a l'evidència del fracàs de l'equip. Farem servir la **fórmula de la probabilitat total condicionada**, que és com (3) però condicionant les probabilitats a una evidència E:

$$\boxed{P(A/E) = \sum_{i=1}^{n} P(A/B_i, E)\, P(B_i/E).} \tag{5}$$

Així tenim:

$$\begin{aligned}
P(\mathrm{W} = \mathrm{Sí}/\mathrm{Z} = \mathrm{No}) &= P(\mathrm{W} = \mathrm{Sí}/\mathrm{Y} = \mathrm{Sí}, \mathrm{T} = \mathrm{Sí}, \mathrm{Z} = \mathrm{No}) P(\mathrm{Y} = \mathrm{Sí}, \mathrm{T} = \mathrm{Sí}/\mathrm{Z} = \mathrm{No}) \\
&+ P(\mathrm{W} = \mathrm{Sí}/\mathrm{Y} = \mathrm{Sí}, \mathrm{T} = \mathrm{No}, \mathrm{Z} = \mathrm{No}) P(\mathrm{Y} = \mathrm{Sí}, \mathrm{T} = \mathrm{No}/\mathrm{Z} = \mathrm{No}) \\
&+ P(\mathrm{W} = \mathrm{Sí}/\mathrm{Y} = \mathrm{No}, \mathrm{T} = \mathrm{Sí}, \mathrm{Z} = \mathrm{No}) P(\mathrm{Y} = \mathrm{No}, \mathrm{T} = \mathrm{Sí}/\mathrm{Z} = \mathrm{No}) \\
&+ P(\mathrm{W} = \mathrm{Sí}/\mathrm{Y} = \mathrm{No}, \mathrm{T} = \mathrm{No}, \mathrm{Z} = \mathrm{No}) P(\mathrm{Y} = \mathrm{No}, \mathrm{T} = \mathrm{No}/\mathrm{Z} = \mathrm{No}) \\
&= 0{,}6 \times P(\mathrm{Y} = \mathrm{Sí}/\mathrm{Z} = \mathrm{No}) \times 0{,}3 + 1{,}0 \times P(\mathrm{Y} = \mathrm{Sí}/\mathrm{Z} = \mathrm{No}) \times 0{,}7 \\
&+ 0{,}0 \times P(\mathrm{Y} = \mathrm{No}/\mathrm{Z} = \mathrm{No}) \times 0{,}3 + 0{,}0 \times P(\mathrm{Y} = \mathrm{No}/\mathrm{Z} = \mathrm{No}) \times 0{,}7 \\
&= 0{,}88 \times P(\mathrm{Y} = \mathrm{Sí}/\mathrm{Z} = \mathrm{No})
\end{aligned} \tag{6}$$

(compareu amb (4)). Noteu que hem hagut de calcular les probabilitats conjuntes de les variables Y: Col·lisió amb la Terra i T: Refugi subterrani condicionades a l'evidència E: Z $=$ No, per exemple $P(\mathrm{Y} = \mathrm{Sí}, \mathrm{T} = \mathrm{Sí}/\mathrm{Z} = \mathrm{No})$ (i el mateix amb les altres tres probabilitats). De manera anàloga al que hem fet quan no teníem cap evidència, ara tenim que

$$P(\mathrm{Y}=\mathrm{Sí},\mathrm{T}=\mathrm{Sí}/\mathrm{Z}=\mathrm{No})=P(\mathrm{Y}=\mathrm{Sí}/\mathrm{Z}=\mathrm{No})P(\mathrm{T}=\mathrm{Sí})=P(\mathrm{Y}=\mathrm{Sí}/\mathrm{Z}=\mathrm{No})\times 0{,}3. \qquad (7)$$

Aquí hem fet servir de nou (5) condicionant als possibles valors, però ara només del pare X: Trajectòria de col·lisió de Y: Col·lisió amb la Terra, ja que el valor de l'altre pare, Z: Explosió meteorit, està fixat perquè estem condicionant a l'evidència E: Z $=$ No. D'aquesta manera hem obtingut que

$$\begin{aligned}P(\mathrm{Y}=\mathrm{Sí},\mathrm{T}=\mathrm{Sí}/\mathrm{Z}=\mathrm{No}) &= P(\mathrm{Y}=\mathrm{Sí},\mathrm{T}=\mathrm{Sí}/\mathrm{X}=\mathrm{Sí},\mathrm{Z}=\mathrm{No})\,P(\mathrm{X}=\mathrm{Sí}/\mathrm{Z}=\mathrm{No})\\ &\quad + P(\mathrm{Y}=\mathrm{Sí},\mathrm{T}=\mathrm{Sí}/\mathrm{X}=\mathrm{No},\mathrm{Z}=\mathrm{No})\,P(\mathrm{X}=\mathrm{No}/\mathrm{Z}=\mathrm{No}),\end{aligned}$$

que és igual a

$$\begin{aligned}&P(\mathrm{Y}=\mathrm{Sí}/\mathrm{X}=\mathrm{Sí},\mathrm{Z}=\mathrm{No})\,P(\mathrm{T}=\mathrm{Sí}/\mathrm{X}=\mathrm{Sí},\mathrm{Z}=\mathrm{No})\,P(\mathrm{X}=\mathrm{Sí})\\ &+P(\mathrm{Y}=\mathrm{Sí}/\mathrm{X}=\mathrm{No},\mathrm{Z}=\mathrm{No})\,P(\mathrm{T}=\mathrm{Sí}/\mathrm{X}=\mathrm{No},\mathrm{Z}=\mathrm{No})\,P(\mathrm{X}=\mathrm{No})\end{aligned}$$

usant la condició de Markov (Y i T són independents donades X i Z, i també X és independent de Z). Si ara usem que T és independent de X i Z (de nou per la condició de Markov), tenim finalment (7):

$$\begin{aligned}P(\mathrm{Y}=\mathrm{Sí},\mathrm{T}=\mathrm{Sí}/\mathrm{Z}=\mathrm{No}) &= P(\mathrm{Y}=\mathrm{Sí}/\mathrm{X}=\mathrm{Sí},\mathrm{Z}=\mathrm{No})\,P(\mathrm{T}=\mathrm{Sí})\,P(\mathrm{X}=\mathrm{Sí})\\ + P(\mathrm{Y}=\mathrm{Sí}/\mathrm{X}=\mathrm{No},\mathrm{Z}=\mathrm{No})\,P(\mathrm{T}=\mathrm{Sí})\,P(\mathrm{X}=\mathrm{No}) &= P(\mathrm{Y}=\mathrm{Sí}/\mathrm{Z}=\mathrm{No})\,P(\mathrm{T}=\mathrm{Sí}),\end{aligned}$$

on en la darrera igualtat hem tret factor comú de $P(\mathrm{T}=\mathrm{Sí})$ i hem fet servir la fórmula de la probabilitat total condicionada.

Per obtenir el valor numèric final a (6) necessitem calcular $P(\mathrm{Y}=\mathrm{Sí}/\mathrm{Z}=\mathrm{No})$, i això ho fem, de nou, amb la fórmula de la probabilitat total:

$$\begin{aligned}P(\mathrm{Y}=\mathrm{Sí}/\mathrm{Z}=\mathrm{No}) &= P(\mathrm{Y}=\mathrm{Sí}/\mathrm{X}=\mathrm{Sí},\mathrm{Z}=\mathrm{No})P(\mathrm{X}=\mathrm{Sí}/\mathrm{Z}=\mathrm{No})\\ &\quad + P(\mathrm{Y}=\mathrm{Sí}/\mathrm{X}=\mathrm{No},\mathrm{Z}=\mathrm{No})P(\mathrm{X}=\mathrm{No}/\mathrm{Z}=\mathrm{No})\\ &= P(\mathrm{Y}=\mathrm{Sí}/\mathrm{X}=\mathrm{Sí},\mathrm{Z}=\mathrm{No})P(\mathrm{X}=\mathrm{Sí})\\ &\quad + P(\mathrm{Y}=\mathrm{Sí}/\mathrm{X}=\mathrm{No},\mathrm{Z}=\mathrm{No})P(\mathrm{X}=\mathrm{No})\\ &= 1{,}0\times 0{,}999+0{,}0\times 0{,}001=0{,}999\end{aligned}$$

Finalment, el risc *a posteriori* donada l'evidència que l'expedició per fer explotar el meteorit ha fracassat, s'obté susbstituint aquest valor a (6):

$$P(\mathrm{W}=\mathrm{Sí}/\mathrm{Z}=\mathrm{No})=0{,}88\times 0{,}999=0{,}87912\approx 88\%.$$

De manera anàloga, es podria obtenir el risc *a posteriori* donada l'evidència que l'expedició per fer explotar el meteorit ha tingut èxit, si fos el cas, i donaria el valor següent:

$$P(\mathrm{W}=\mathrm{Sí}/\mathrm{Z}=\mathrm{Sí})=0{,}88\times 0{,}1998=0{,}175824\approx 18\%.$$

Hem vist que, com és lògic, fracassar en l'expedició fa augmentar el risc (el multiplica aproximadament per **1,1**), mentre que tenir èxit el fa disminuir (el multiplica per **0,22** o, el que és el mateix, el divideix per **4,5**). El model, però, ens permet quantificar tant l'augment com la disminució. I, a més, ens permet calcular fàcilment el risc *a posteriori* en escenaris més complexos, en els quals l'evidència que tenim involucra més d'una variable, quan la intuïció deixa de poder ajudar-nos. Per exemple, podem considerar escenaris en els quals tinguem evidència sobre l'èxit o el fracàs tant de l'expedició al meteorit per fer-lo explotar, com de la mesura mitigant consistent a preparar refugis subterranis per a la població. La taula de la figura 11 mostra que sempre que tinguem èxit en alguna de les dues mesures, el risc disminueix, però que si hem de triar, és més rendible invertir en l'expedició al meteorit que en la mesura mitigant, encara que el millor, lògicament, és que les dues tinguin èxit, en aquest cas obtindrem una reducció màxima del risc fins al 12%.

Estimació del risc d'*Armaggeddon*			
A priori	**81%**		
A posteriori	Evidència Explosió = Sí, Refugi = No	Evidència Explosió = No, Refugi = Sí	Evidència Explosió = Sí, Refugi = Sí
	20%	**60%**	**12%**
	x 0,25	**x 0,74**	**x 0,15**

Figura 11. Risc *a posteriori* en escenaris més complexos de l'exemple de l'*Armaggeddon*.

En aquest exemple hem fet servir la fórmula de la probabilitat total condicionada per actualitzar riscos, és a dir, per calcular el risc *a posteriori* a partir d'una evidència. Depenent de com sigui la xarxa bayesiana, en comptes d'aplicar aquesta fórmula se n'haurà d'aplicar una altra, la fórmula de Bayes. Per explicar què és i com es pot fer servir, introduïm un exemple diferent en la secció següent.

2.7. Un exemple mèdic: el risc de patir càncer de pulmó

Volem avaluar el risc de patir càncer de pulmó d'un individu. Per fer-ho, construïm un model amb la variable Càncer (Sí/No) i altres variables que poden ser rellevants per al diagnòstic:

Variables sociodemogràfiques

- Contaminació: indica si la persona viu en una zona amb contaminació ambiental alta o no (Alta/Baixa). L'abreugem i queda Contam.
- Fumador: indica si la persona és o no fumadora (Sí/No). L'abreugem i queda Fuma.

Variables clíniques

- Raigs X: resultat de la prova diagnòstica de raigs X (positiu +/negatiu −).
- Dispnea: indica si la persona pateix aquesta dificultat respiratòria o no (Sí/No). L'abreugem i queda Disp.

El model que construïm és la xarxa bayesiana que ve donada pel DAG i les probabilitats condicionades de la figura 12.

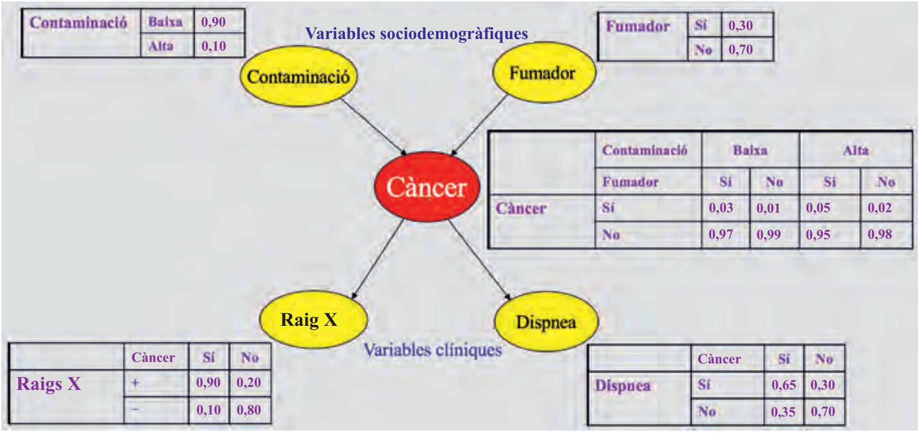


Figura 12. Xarxa bayesiana per a l'exemple del càncer de pulmó.

Calculem en primer lloc el risc *a priori* de patir càncer de pulmó $P(\text{Càncer} = \text{Sí})$ fent servir la fórmula de la probabilitat total, condicionant pels valors dels dos pares de la variable Càncer, que són Contaminació i Fumador.

$$
\begin{aligned}
&P(\text{Càncer} = \text{Sí})\\
&= P(\text{Càncer} = \text{Sí}/\text{Contam} = \text{Baixa,Fuma} = \text{Sí})P(\text{Contam} = \text{Baixa,Fuma} = \text{Sí})\\
&+ P(\text{Càncer} = \text{Sí}/\text{Contam} = \text{Baixa, Fuma} = \text{No})\, P(\text{Contam} = \text{Baixa, Fuma} = \text{No})\\
&+ P(\text{Càncer} = \text{Sí}/\text{Contam} = \text{Alta,Fuma} = \text{Sí})P(\text{Contam} = \text{Alta,Fuma} = \text{Sí})\\
&+ P(\text{Càncer} = \text{Sí}/\text{Contam} = \text{Alta, Fuma} = \text{No})\, P(\text{Contam} = \text{Alta, Fuma} = \text{No})\\
&= 0{,}03 \times 0{,}90 \times 0{,}30 + 0{,}01 \times 0{,}90 \times 0{,}70 + 0{,}05 \times 0{,}10 \times 0{,}30 + 0{,}02 \times 0{,}10 \times 0{,}70\\
&= 0{,}0173\,,
\end{aligned}
$$

on hem fet servir que les variables Contaminació i Fumador són independents i, per tant,

$$P(\text{Contam} = \text{Baixa,Fuma} = \text{Sí}) = P(\text{Contam} = \text{Baixa})\, P(\text{Fuma} = \text{Sí})\,,$$

i anàlogament per a les altres tres probabilitats conjuntes de les variables Contaminació i Fumador. Per tant, el risc *a priori* de patir càncer de pulmó ha resultat:

$$P(\text{Càncer} = \text{Sí}) = 0{,}0173 \approx 17\,‰$$

Avaluant el risc *a posteriori*: la fórmula de Bayes

Com es modifica el risc *a priori* si sabem que una persona pateix de dispnea, per exemple? Intuïtivament, el risc augmentarà, però quant? Per contestar a la pregunta hem de calcular la probabilitat de patir càncer si conexíem l'evidència que la persona té dispnea, és a dir, hem de calcular $P(\text{Càncer} = \text{Sí}/\text{Dispnea} = \text{Sí})$, però no ho podem fer mitjançant la fórmula de la probabilitat total condicionada, com hem fet amb l'exemple de l'*Armaggeddon*, ja que la variable per la qual condicionem no és un *ancestre*, sinó un *descendent* (de fet, és un *fill*) de la variable Càncer. Així doncs, hem de calcular una probabilitat condicionada en un sentit invers de com venen donades les probabilitats per la xarxa bayesiana (probabilitats de la figura

12). Per fer-ho, ens ajudarem de la fórmula (2). En efecte, fent-la servir amb A: Càncer = Sí i B_i: Dispnea = Sí, podem aïllar la probabilitat condicionada i tenim que:

$$P(\text{Càncer} = \text{Sí}/\text{Dispnea} = \text{Sí}) = \frac{P(\text{Càncer} = \text{Sí} \cap \text{Dispnea} = \text{Sí})}{P(\text{Dispnea} = \text{Sí})}. \tag{8}$$

I ara l'apliquem de nou al numerador de (8) per expressar la probabilitat de la intersecció en termes de la probabilitat que coneixem, que és la de Dispnea = Sí condicionada a Càncer = Sí. En efecte, tenim que

$$P(\text{Càncer} = \text{Sí} \cap \text{Dispnea} = \text{Sí}) = P(\text{Dispnea} = \text{Sí}/\text{Càncer} = \text{Sí})\, P(\text{Càncer} = \text{Sí})$$

i substituint a (8) tenim:

$$P(\text{Càncer} = \text{Sí}/\text{Dispnea} = \text{Sí}) = \frac{P(\text{Dispnea} = \text{Sí}/\text{Càncer} = \text{Sí})\, P(\text{Càncer} = \text{Sí})}{P(\text{Dispnea} = \text{Sí})}. \tag{9}$$

En general, la fórmula (9) es pot expressar així:

$$\boxed{P(A/B) = \frac{P(B/A)\, P(A)}{P(B)}} \tag{10}$$

i es coneix com a **fórmula de Bayes**. Tal com ja hem comentat en la introducció, va ser introduïda per Thomas Bayes el 1763 a [2].

Apliquem, doncs, la fórmula de Bayes (9) amb les probabilitats de la figura 12, tenint en compte que $P(\text{Càncer} = \text{Sí}) = 0{,}0173$ ho hem calculat abans (és el risc *a priori*), i que per obtenir el denominador $P(\text{Dispnea} = \text{Sí})$ hem de fer servir la fórmula de la probabilitat total (3). Així, obtenim:

$$\begin{aligned}
&P(\text{Càncer} = \text{Sí}/\text{Dispnea} = \text{Sí}) = \\
&\frac{P(\text{Dispnea} = \text{Sí}/\text{Càncer} = \text{Sí})\, P(\text{Càncer} = \text{Sí})}{P(\text{Disp} = \text{Sí}/\text{Càncer} = \text{Sí})P(\text{Càncer} = \text{Sí}) + P(\text{Disp} = \text{Sí}/\text{Càncer} = \text{No})P(\text{Càncer} = \text{No})} \\
&= \frac{0{,}65 \times 0{,}0173}{0{,}65 \times 0{,}0173 + 0{,}30 \times (1 - 0{,}0173)} = 0{,}03674 \approx 37‰
\end{aligned}$$

Ara tenim una nova evidència sobre la persona que sabem que pateix dispnea: se li ha fet la prova de raigs X i ha donat postitiu (+). Com augmentarà això el risc? Hem de fer el càlcul següent, aplicant de nou la fórmula de Bayes (i, al denominador, la fórmula de la probabilitat total), per calcular la probabilitat de Càncer = Sí condicionada a l'esdeveniment Dispnea = Sí, Raigs X = +, que és una intersecció de dos esdeveniments elementals.

$$\begin{aligned}
&P(\text{Càncer} = \text{Sí}/\text{Dispnea} = \text{Sí, Raigs X} = +) = \\
&\frac{P(\text{Dispnea} = \text{Sí, Raigs X} = +/\text{Càncer} = \text{Sí})\, P(\text{Càncer} = \text{Sí})}{P(\text{Dispnea} = \text{Sí,Raigs X} = +/\text{Càncer} = \text{Sí})P(\text{Càncer} = \text{Sí}) + P(\text{Dispnea} = \text{Sí,Raigs X} = +/\text{Càncer} = \text{No})P(\text{Càncer} = \text{No})} \\
&= \frac{0{,}65 \times 0{,}90 \times 0{,}0173}{0{,}65 \times 0{,}90 \times 0{,}0173 + 0{,}30 \times 0{,}20 \times (1 - 0{,}0173)} = 0{,}146499 \approx 146‰
\end{aligned}$$

Noteu que, a causa de la condició de Markov, les variables Dispnea i Raigs X són **independents** atesa la variable pare Càncer (vegeu la figura 12). Per això hem pogut calcular $P(\text{Dispnea} = \text{Sí}, \text{Raigs X} = +/\text{Càncer} = \text{Sí})$ com el producte de probabilitats $P(\text{Dispnea} = \text{Sí}/\text{Càncer} = \text{Sí})\, P(\text{Raigs X} = +/\text{Càncer} = \text{Sí}) = 0{,}65 \times 0{,}90$, i anàlogament quan es condiciona per Càncer = No.

Les estimacions dels riscos que hem obtingut fins ara en l'exemple del càncer de pulmó apareixen a la figura 13, on veiem que tenir només l'evidència de patir dispnea multiplica per 2,18 el risc, mentre que si afegim un resultat positiu en la prova de raigs X, el risc es multiplica per 8,6. És a dir, tant patir dispnea com un resultat positiu en la prova de raigs X, són **factors de risc** per al càncer de pulmó.

Estimació del risc de patir càncer de pulmó		
A priori	17 ‰	
A posteriori	Evidència Dispnea = Sí	Evidència Dispnea = Sí Raigs X = +
	37 ‰	146 ‰
	↗ x 2,18	↗ x 8,6

Figura 13. Riscos *a priori* i *a posteriori* per a l'exemple del càncer de pulmó.

I què podem dir sobre el tabac? Ser fumador és un factor de risc? Vegem-ho. Calcularem $P(\text{Càncer} = \text{Sí}/\text{Fumador} = \text{Sí})$ fent servir la fórmula de la probabilitat total condicionada (5) amb l'evidència E: Fumador = Sí. L'esdeveniment per al qual volem calcular la probabilitat és A: Càncer = Sí, i la resta són B_1: Contaminació = Baixa i B_2: Contaminació = Alta. Així, tenim que

$$
\begin{aligned}
&P(\text{Càncer} = \text{Sí}/\text{Fumador} = \text{Sí})\\
&= P(\text{Càncer} = \text{Sí}/\text{Contam} = \text{Baixa}, \text{Fuma} = \text{Sí})\, P(\text{Contam} = \text{Baixa}/\text{Fuma} = \text{Sí})\\
&+ P(\text{Càncer} = \text{Sí}/\text{Contam} = \text{Alta}, \text{Fuma} = \text{Sí})\, P(\text{Contam} = \text{Alta}/\text{Fuma} = \text{Sí})\\
&= 0{,}03 \times 0{,}90 + 0{,}05 \times 0{,}10 = 0{,}032 = 32‰
\end{aligned}
$$

on hem fet servir que, per la condició de Markov, les variables Contaminació i Fumador són **independents** i, per tant, $P(\text{Contam} = \text{Baixa}/\text{Fuma} = \text{Sí}) = P(\text{Contam} = \text{Baixa})$ i $P(\text{Contam} = \text{Alta}/\text{Fuma} = \text{Sí}) = P(\text{Contam} = \text{Alta})$.

Anàlogament, s'obté que $P(\text{Càncer} = \text{Sí}/\text{Fumador} = \text{No}) = 0{,}01 \times 0{,}90 + 0{,}02 \times 0{,}10 = 0{,}011 = 11‰$. És a dir, pel que fa al càncer de pulmó, fumar és un **factor de risc**. Els resultats queden resumits en la taula de la figura 14.

Estimació del risc de patir càncer de pulmó		
A priori	17 ‰	
A posteriori	Evidència Fumador = Sí	Evidència Fumador = No
	32 ‰	11 ‰
	↗ x 1,88	↘ x 0,65

Figura 14. El tabaquisme com a factor de risc per al càncer de pulmó.

Observació: en els exemples que hem introduït hem pogut fer els càlculs a mà perquè les xarxes bayesianes eren senzilles, és a dir, amb pocs nodes, i no gaire connectades (poques fletxes). En una situació real, amb molts nodes i connexions, és necesari l'ús d'un algortime informàtic que ens faciliti els càlculs. D'altra banda, en els exemples previs les xarxes bayesianes, tant les seves estructures o DAG, com els seus paràmetres, que són les probabilitats condicionades de cada node als seus pares en el DAG, incloent-hi les probabilitats sense condicionar dels nodes arrels, ens han vingut donades. A la pràctica, però, tant una cosa com l'altra s'han d'aprendre. Encara que l'opinió dels experts pot ser tinguda en compte, el més objectiu i habitual és fer aquest aprenentatge a partir d'una base de dades, mitjançant uns algoritmes informàtics implementats en algun lleguatge de programació, com ara l'R[4], que disposa de llibreries específiques per treballar amb xarxes bayesianes. Per exemple, en el cas del càncer de pulmó, seria a partir d'una base de dades de (molts) casos de persones per a les quals es coneixien les cinc variables de model, és a dir, si vivien en una zona amb contaminació alta o baixa, si eren o no fumadors, si patien de dispnea, si el resultat de la prova de raigs X havia estat positiu o negatiu i, finalment, si tenien o no càncer de pulmó. La base de dades es fa servir per aprendre el model, tant el DAG com els paràmetres, i també per validar-lo. Un cop validat, ja es pot fer servir per estimar les probabilitats d'interès, com ara el risc de partir càncer de pulmó d'un nou individu a partir del coneixement que tinguem sobre les dues variables sociodemogràfiques i les dades clíniques incloses en el model. Els detalls tècnics dels algoritmes que es fan servir per a l'aprenentatge automàtic de la xarxa bayesiana, cosa que justifica que es consideri una eina d'**aprenentatge automàtic** (*Machine Learning* en anglès), queden fora de l'abast d'aquest article.

3. Avaluant evidències: l'odds ratio o raó de contraris

Als països anglosaxons és molt comú fer servir l'**odds** com a mesura del grau d'incertesa d'un esdeveniment, especialment en jocs d'atzar i al món de les apostes. Per exemple, es diu que «un equip de fútbol té un *odds a favor* de 3 a 2 de guanyar un partit». Què vol dir això? Donem primer la definició:

Definició: donat un esdeveniment A, es defineix l'**odds en contra** d'A com $\text{Odds}(A^c) = \frac{P(A^c)}{P(A)}$, i es defineix l'**odds a favor** d'A com $\text{Odds}(A) = \frac{P(A)}{P(A^c)}$, és A^c on el complementari (o contrari) de l'esdeveniment A, de tal manera que $P(A^c) = 1 - P(A)$.

Així, si diem que l'equip té un odds a favor de guanyar de 3 a 2, vol dir que si denotem per A l'esdeveniment «l'equip guanya el partit», $\frac{P(A)}{P(A^c)} = \frac{3}{2}$. En particular, això ens diu que $P(A) > P(A^c)$, és a dir, que la seva probabilitat de guanyar és més gran que la de perdre. Però podem trobar quina és exactament la probabilitat de guanyar el partit a partir de l'odds així:

$$\frac{3}{2} = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)} \Longrightarrow P(A) = \frac{3}{5} = 0{,}6 \tag{11}$$

4. R és GNU S, un llenguatge i entorn lliure disponible per a la computació estadística i gràfica que proporciona una àmplia varietat de tècniques estadístiques i gràfiques: modelització lineal i no lineal, proves estadístiques, anàlisi de sèries temporals, classificació, etc. https://cran.r-project.org.

Per tant, dir que «un equip de fútbol té un *odds a favor* de 3 a 2 de guanyar un partit» és equivalent a dir que «la seva probabilitat de guanyar el partir és $3/5 = 0{,}6$» (és una equivalència perquè en realitat la implicació de (11) és un *si i només si*).

Com que els riscos són probabilitats, podem expressar-los en termes d'odds. Tornem a l'exemple de l'*Armaggeddon*, quan diem que el risc *a priori* és 0,8087904 (aproximadament un 81%), equivalentment estem dient que l'odds a favor de l'*Armaggeddon* és aproximadament 4,23. És a dir, les apostes *a priori* estan aproximadament 4 a 1 a favor de l'*Armaggeddon*. En efecte, si diem A a l'esdeveniment *Armaggeddon* = Sí, tenim que $P(A) = 0{,}8087904$ i, per tant, l'odds a favor de l'esdeveniment A és:

$$\frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)} = \frac{0{,}8087904}{1 - 0{,}8087904} = 4{,}229863 \approx 4{,}23$$

(les apostes estan aproximadament 4 a 1 a favor de l'*Armaggeddon*).

Si ara considerem el risc *a posteriori* condicionat a, per exemple, l'evidència E: Explosió = Sí, Refugi = No, com que la probabilitat de l'esdeveniment A condicionada era $P(A/E) \approx 0{,}20$ (vegeu la taula de la figura 11), tenim que ara l'odds a favor de l'esdeveniment A condicionat a l'evidència E és 0,25. És a dir, que ara les apostes estan 4 a 1 en contra de l'*Armaggeddon*. Naturalment, ha canviat molt!

En el cas del risc *a priori*, obtenim un odds *a priori* a favor de l'*Armaggeddon* de $4{,}23 > 1$, que vol dir que és més probable que succeeixi, mentre que l'odds *a posteriori* a favor és $0{,}25 < 1$, que vol dir que en el nou escenari és més probable que no succeeixi. Si fem el quocient, tenim l'**odds ratio**:

$$\text{Odds ratio a favor} = \frac{\text{odds } \textit{a posteriori} \text{ a favor}}{\text{odds } \textit{a priori} \text{ a favor}} \approx \frac{0{,}25}{4{,}23} = 0{,}0591$$

que ens indica quant han baixat les apostes en favor de l'*Armaggeddon* atesa l'evidència. S'han multiplicat aproximadament per 0,059 o, el que és el mateix, s'han dividit per 16,95.

Podem reescriure els resultats de les taules de riscos ara en termes d'odds *a priori* i *a posteriori*, i de l'odds ratio. Així, la taula de la figura 11 serà equivalent a la de la figura 15.

<table>
<tr><th colspan="4">Odds a favor de l'Armaggeddon</th></tr>
<tr><td>A priori</td><td colspan="3">4,23 (4 a 1, a favor de l'Armaggeddon)</td></tr>
<tr><td rowspan="2">A posteriori</td><td>Evidència
Explosió = Sí, Refugi = No</td><td>Evidència
Explosió = No, Refugi = Sí</td><td>Evidència
Explosió = Sí, Refugi = Sí</td></tr>
<tr><td>0,25 (4 a 1, en contra de l'Armaggeddon)</td><td>1,5 (3 a 2, a favor de l'Armaggeddon)</td><td>0,136 (7 a 1, en contra de l'Armaggeddon)</td></tr>
<tr><td>Odds ratio</td><td>0,059</td><td>0,355</td><td>0,032</td></tr>
</table>

Figura 15. Odds *a priori*, *a posteriori* i odds ratio en l'exemple de l'*Armaggeddon*.

L'odds ratio ens permet avaluar el pes de les evidències. En el cas de les de la taula de la figura 15, l'evidència que més canvia l'odds *a priori* és la de Explosió = Sí, Refugi = Sí, que la fa canviar de 4 a 1 a favor de l'*Armaggeddon*, a 7 a 1 en contra. És la més influent. La segona més influent és Explosió = Sí, Refugi = No, que la fa passar de 4 a 1 a favor, a 4 a 1 en contra. I la menys influent, que només la fa passar de 4 a 1 a favor, a 3 a 2 a favor, és Explosió = No, Refugi = Sí.

Anàlogament, a partir de la taula de riscos de les figures 13 i 14, tenim la taula següent, en la figura 16, amb l'odds *a priori*, *a posteriori*, i l'odds ratio en l'exemple del càncer de pulmó.

Odds en contra de patir càncer de pulmó				
A priori	0,0176 (57 a 1, en contra del càncer)			
	Evidència Dispnea = Sí	Evidència Dispnea = Sí Raigs X = +	Evidència Fumador = Sí	Evidència Fumador = No
A posteriori	0,0367 (27 a 1, en contra del càncer)	0,1465 (7 a 1, en contra del càncer)	0,032 (31 a 1, en contra del càncer)	0,011 (91 a 1, en contra del càncer)
Odds ratio	2,09	8,32	1,82	0,625

Figura 16. Odds *a priori*, *a posteriori* i odds ratio en l'exemple del càncer de pulmó.

Aquí veiem que l'evidència de no ser fumador fa passar d'un 57 a 1 en contra de patir càncer de pulmó, a un 91 a 1 en contra. En canvi, les altres evidències considerades són factors de riscs, ja que es passa a estar en contra de patir càncer de 31 a 1, de 27 a 1 o de 7 a 1 (en ordre de menys a més importància de l'evidència). Per exemple, podem veure que ser fumador és un factor de risc, però una mica menys que patir de dispnea, mentre que patir de dispnea i donar positiu en la prova de raigs X és el factor de risc més important dels que hem considerat en l'exemple.

Interpretació de l'odds ratio a través de la raó de versemblances

Considerem, en general, que tenim un esdeveniment A i l'odds a favor d'A: $\text{Odds}(A) = \frac{P(A)}{P(A^c)}$, que és l'odds a priori. Donada una evidència E, tenim l'odds *a posteriori* a favor d'A: $\text{Odds}(A/E) = \frac{P(A/E)}{P(A^c/E)}$. Com es relacionen? Recordem que per la Fórmula de Bayes (10),

$$P(A/E) = \frac{P(E/A)\,P(A)}{P(E)} \quad \text{i} \quad P(A^c/E) = \frac{P(E/A^c)\,P(A^c)}{P(E)}.$$

Per tant, si fem el quocient, tindrem:

$$\frac{P(A/E)}{P(A^c/E)} = \frac{P(E/A)}{P(E/A^c)}\,\frac{P(A)}{P(A^c)},$$

ja que $P(E)$ se simplifica i com a conseqüència tenim que

$$\mathrm{Odds}(A/E) = \frac{P(E/A)}{P(E/A^c)}\,\mathrm{Odds}(A)\,. \tag{12}$$

El quocient $\frac{P(E/A)}{P(E/A^c)}$ s'anomena **raó de versemblances** (*likelihood ratio* en anglès, **LR**) perquè s'obté dividint $P(E/A)$, que és una mesura de com de versemblant és l'evidència si sabem que l'esdeveniment A s'ha produït, entre $P(E/A^c)$, que mesura com de versemblant és l'evidència si sabem que l'esdeveniment A no s'ha produït. Per tant, aïllant de (12) tenim que la *raó de versemblances* és l'odds ratio a favor d'A:

$$\boxed{\mathbf{LR} = \frac{\mathrm{Odds}(A/E)}{\mathrm{Odds}(A)}}$$

La interpretació de la *raó de versemblances* és com segueix, en línies molt generals:

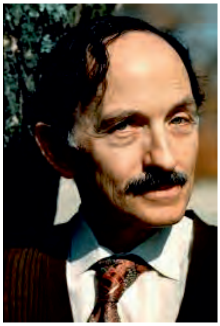
$$\mathbf{LR}\begin{cases} \approx 1 & \text{l'evidència E no dona suport a A ni al seu contrari,}\\ > 1 & \text{l'evidència E dona suport a A (més, com més gran sigui),}\\ < 1 & \text{l'evidència E dona suport al contrari d'A (més, com més petit sigui),}\end{cases}$$

encara que en determinats àmbits s'ha concretat més. Per exemple, en ciències forenses s'usa l'**escala SKL** (*Swedish National Laboratory of Forensic Science*) [14].

Una mica d'història sobre la raó de versemblances

A la dècada del 1940 el matemàtic britànic Alan Turing i el seu equip van fer servir l'LR per desxifrar els codis de la màquina Enigma dels nazis a Bletchley Park. Irving John *Jack* Good, un altre matemàtic britànic, company de Turing, va difondre'n lús en la ciència en general amb el seu llibre *Probability and the weighing of evidence*, publicat el 1950 [8]. Com a dada curiosa sobre Good, direm que Stanley Kubrick, director de *2001, una odisea a l'espai* (1968), el va consultar com a especialista en supercomputadors per a la seva pel·lícula.

(a) A. Turing (Paddington, Londres, 1912 — Wilmslow, Cheshire, 1954).



(b) I. J. Good (Londres, 1916 — Radford, Virginia, 2009).

Figura 17

Més tard, Dennis Lindley, estadístic britànic, que va ser un dels fundadors de l'escola d'estadística bayesiana, va contribuir a la introducció de l'LR de manera rigorosa en la ciència forense, a partir d'evidències recollides a l'escenari d'un delicte relatives a fragments de vidre, en un article del 1977 [11]. A partir de la dècada del 1990 es va començar a fer servir l'LR per ajudar a interpretar evidències d'ADN, especialment en tribunals de justícia dels països anglosaxons.

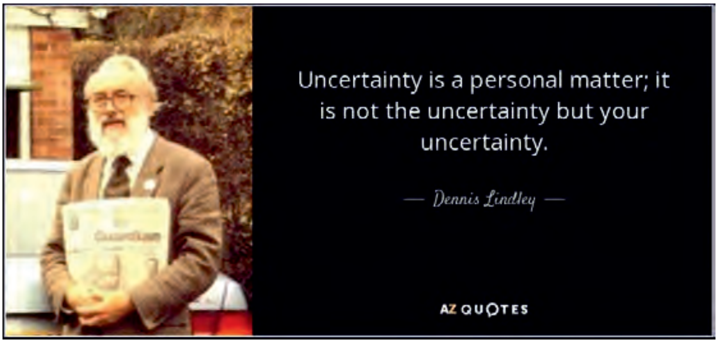


Figura 18. Dennis Lindley (Surbiton, Londres, 1923 — Somerset, 2013).

Un exemple final: el cas de l'atracador disfressat

Un home alt, de 180 cm d'alçada, és detingut com a sospitós d'haver atracat un banc disfressat. La policia ha pogut accedir a una imatge digital gravada de l'atracador i ha pogut deduir que la seva alçada és de més de 175 cm. Per tant, tenim l'evidència

E: L'atracador és una persona alta.

Com afecta aquesta evidència a l'odds en favor que el detingut és culpable?

Diem A a l'esdeveniment El detingut és culpable. Per tant, el seu complementari o contrari és A^c: El detingut és innocent. Per obtenir l'odds *a priori* en favor de la culpabilitat del detingut, necessitem poder calcular $P(A)$, que pot dependre d'altres circumstàncies del cas que no coneixem. No importa! Fent servir la raó de versemblances sabrem per quant multiplica l'odds *a priori* l'evidència, independentment de quin sigui el seu valor. En efecte, calcularem:

$$\mathbf{LR} = \frac{P(E/A)}{P(E/A^c)}$$

i per fer-ho tenim en compte que $P(E/A) = 1$, ja que si el sospitós és culpable, com que és alt, segur que en observar l'alçada de l'atracador veurem que és una persona alta. D'altra banda, per calcular la probabilitat del denominador, de quina informació disposem? La representarem mitjançant una xarxa bayesiana amb tres nodes: Culpable (Sí/No) (referint-se al sospitós), Sexe (Home/Dona) i Alçada (Alt/Baix), ambdues referides a l'atracador, tal com apareix en la figura 19. Per tant, amb les notacions introduïdes, A: Culpable = Sí i E: Alçada = Alt.

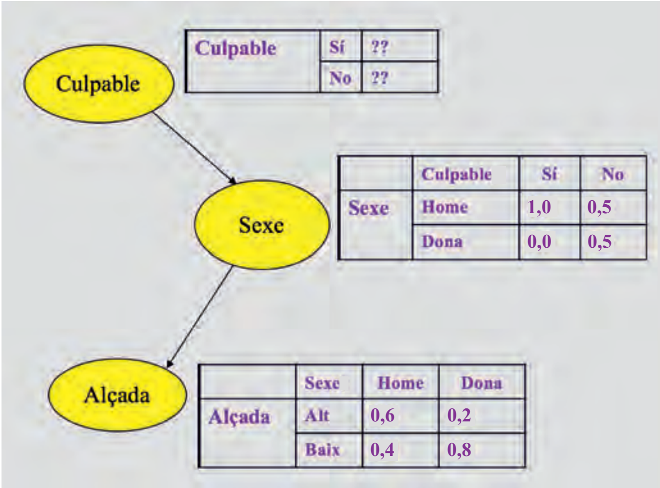


Figura 19. Xarxa bayesiana de l'exemple final de l'atracador disfressat.

En aquest model hem inclòs les dades que coneixem, que són que el 50% de la població són homes (simplificant) i que el 60% dels homes són alts (alçada $>$ 175 cm), mentre que de les dones, només ho són el 20%. Amb aquesta informació, relativament senzilla de trobar, podem calcular la probabilitat que ens interessa fent servir la fórmula de la probabilitat total condicionada (5), condicionant als esdeveniments B_1: Sexe = Home i B_2: Sexe = Dona, amb E: Alçada = Alt i A^c : Culpable = No, així:

$$\begin{aligned} P(E/A^c) &= P(E/Home, A^c)\, P(Home/A^c) + P(E/Dona, A^c)\, P(Dona/A^c) \\ &= 0{,}6 \times 0{,}5 + 0{,}2 \times 0{,}5 = 0{,}4 \end{aligned}$$

Noteu que per fer aquest càlcul no hem necessitat conèixer la probabilitat *a priori* de culpabilitat del sospitós i que per la condició de Markov, donada la informació sobre el sexe de l'individu, E i A^c són independents, per la qual cosa

$$P(E/Home, A^c) = P(E/Home) \quad \text{i} \quad P(E/Dona, A^c) = P(E/Dona)\,.$$

Llavors,

$$\mathbf{LR} = \frac{P(E/A)}{P(E/A^c)} = \frac{1}{0{,}4} = 2{,}5$$

i, per tant, tenim que **sense dependre de l'odds *a priori* en favor de la culpabilitat del sospitós, el fet de conèixer l'evidència que l'atracador és alt, la multiplica per 2,5**. Aquest és el pes de l'evidència!

Bibliografia

[1] Adusei-Poku, K. (2005). *Operational Risk management - Implementing a BN for Foreign Exchange and Money Market Settlement*. Tesi doctoral. Göttingen: University.

[2] Bayes, T. (1763). «An Essay towards solving a Problem in the Doctrine of Chances». *Philosophical Transactions of the Royal Society*, 53, 370-418.

[3] Borsuk, M.E.; Stow, C.A.; Reckhow, K.H. (2004). «A BN of eutrophication models for synthesis, prediction, and uncertainty analysis», *Ecological Modeling*, 173, 219-239.

[4] Cruz-Ramírez, N.; Acosta-Mesa, H.G.; Carrillo-Calvet, H.; Alonso Nava-Fernández, L.; Barrientos-Martínez, R.E. (2007). «Diagnosis of breast cancer using BN: A case study», *Computers in Biology and Medicine*, 37, 1553-1564.

[5] Delgado, R.; Tibau, X-A. (2015). «Las Redes Bayesianas como herramienta para la evaluación del riesgo de reincidencia: Un estudio sobre agresores sexuales», *Revista Española de Investigación Criminológica*, 13.

[6] Delgado, R.; González, J.L.; Sotoca, A.; Tibau, X.A.(2016). «A Bayesian Network Profiler for Wildfire Arsonists». A: Pardalos, P.; Conca, P.; Giuffrida, G.; Nicosia, G. (eds). *Machine Learning, Optimization, and Big Data. MOD 2016. Lecture Notes in Computer Science*, Cham: Springer.

[7] Gagniuc, P.A. (2017). *Markov Chains: From Theory to Implementation and Experimentation*. New Jersey: John Wiley and Sons.

[8] Good, I.J. (1950). *Probability and the weighing of evidence*, 1a ed. C. Griffin.

[9] Korb, K.B.; Nicholson, A.E. (2011). *Bayesian Artificial Intelligence*, 2a ed. CRC Press (Taylor & Francis Group).

[10] Lee, C.; Lee, K.J. (2006). «Application of BN to the probabilistic risk assessment of nuclear waste disposal», *Reliability Engineering and System Safety*, 91(5), 515-532.

[11] Lindley, D.V. (1977). «A problem in forensic science». *Biometrika*, 64(2), 207-213.

[12] Neapolitan, R.E. (1989). *Probabilistic reasoning in expert systems: theory and algorithms*. Wiley.

[13] Neapolitan, R.E. (2004). *Learning Bayesian Networks*. Prentice Hall.

[14] Nordgoaard, A.; Ansell, R.; Drotz, W.; Jaeger, L. (2012). «Scale of conclusions for the value of evidence». *Law, Probability and Risk*, 11, 1-24.

[15] Pearl, J. (1985). Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning. UCLA Technical Report CSD-850017. Proceedings of the 7th Conference of the Cognitive Science Society, Irvine, CA: University of California.

[16] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Francisco CA: Morgan Kaufmann.

[17] Pollino, C.A.; Woodberry, O.; Nicholson, A.; Korb, K.; Hart, B.T. (2007). «Parameterisation and evaluation of a BN for use in an ecological risk assessment», *Environmental Modelling and Software*, 22, 1140-1152.

[18] Spiegelhalter, D.J. (2004). «Incorporating Bayesian ideas into healthcare evaluation», *Statistical Science*, 19, 156-174.

[19] Ticehurst, J.L., Newham, L.T.H., Rissik, D., Letcher, R.A., Jakeman, A.J. (2007). A BN approach for assessing the sustainability of coastal lakes in New South Wales, Australia, Environmental Modelling and Software 22 (8), 1129-1139.

[20] Walshe, T.; Burgman, M. (2010). «A Framework for Assessing and Managing Risks Posed by Emerging Diseases», *Risk Analysis*, 30(2), 236-249.