Butlletí de la Societat Catalana de Matemàtiques
Vol. 17, núm. 1, 2002. Pàg. 7–64

Una introducció a l'estadística bayesiana*

José M. Bernardo

1 Introducció

Els resultats científics o experimentals consisteixen generalment en conjunts de dades de la forma $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, on les $\boldsymbol{x}_i$ són observacions «homogènies» (possiblement multidimensionals). Els mètodes estadístics s'utilitzen per a treure conclusions sobre la naturalesa i el comportament futur del procés que ha produït les observacions. Un element central de *tota* anàlisi estadística és l'especificació d'un *model probabilístic* que se suposa que descriu el mecanisme que ha generat les dades observades D com una funció d'un paràmetre (possiblement multidimensional) $\boldsymbol{\omega} \in \Omega$, sovint anomenat l'*estat de la natura*, sobre el valor del qual només es disposa (en el millor cas) d'informació limitada. És obvi que totes les conclusions estadístiques estan condicionades pel model probabilístic.

A diferència de moltes altres branques de les matemàtiques, els mètodes convencionals d'inferència estadística pateixen de la manca d'una base axiomàtica; en conseqüència, sovint els seus objectius són mútuament incompatibles i l'anàlisi de les mateixes dades pot conduir a resultats contradictoris quan s'utilitzen mètodes aparentment intuïtius però diferents. En contrast, l'aproximació bayesiana a la inferència estadística està fermament basada en fonaments axiomàtics que proporcionen una estructura lògica unificada i garanteixen la consistència mútua dels mètodes proposats. Els mètodes bayesians constitueixen un paradigma *complet* per a la inferència estadística, una revolució científica en el sentit de Kuhn.

*Aquest text és una versió abreujada de l'article *Bayesian Statistics* de l'*Encyclopedia of Life Support Systems* (París: UNESCO), i va servir de base per a la conferència pronunciada a la Quarta Trobada Matemàtica de la Societat Catalana de Matemàtiques, que va tenir lloc a la seu de l'Escola Universitària Politècnica de Vilanova i la Geltrú l'abril de 2001. Aquest treball ha estat parcialment subvencionat pel projecte PB97-1403 de la DGICYT, Madrid.

L'estadística bayesiana només necessita les *matemàtiques* de la teoria de la probabilitat i la *interpretació* de la probabilitat que correspon a l'ús habitual més proper d'aquesta paraula en el llenguatge ordinari; no és per casualitat que alguns dels llibres més importants d'estadística bayesiana, com els treballs de Laplace, de Finetti o Jeffreys, es titulin en realitat *Teoria de la probabilitat*. Les conseqüències pràctiques d'adoptar el paradigma bayesià són de llarg abast. De fet, els mètodes bayesians, *a*) redueixen la inferència estadística a problemes de teoria de la probabilitat i, per tant, minimitzen la necessitat de conceptes completament nous, i *b*) serveixen per a discriminar entre tècniques estadístiques convencionals, tot proporcionant una justificació lògica a algunes (i fent explícites les condicions sota les quals són correctes) o demostrant les inconsistències lògiques de les altres.

La conseqüència més important d'aquesta fonamentació és la *necessitat* matemàtica de descriure totes les incerteses presents en el problema mitjançant distribucions de probabilitat. En particular, els paràmetres desconeguts en els models probabilístics *han* de tenir una distribució de probabilitat conjunta que descrigui tota la informació disponible sobre els seus valors; això es veu sovint com l'element més característic de l'aproximació bayesiana. Cal notar que (a diferència de l'estadística convencional) els *paràmetres es tracten com a variables aleatòries* dintre del paradigma bayesià. Això no és una descripció de la seva variabilitat (els paràmetres són típicament quantitats *fixes desconegudes*) sinó una descripció de la *incertesa* sobre els seus valors autèntics.

Un cas particularment important apareix quan no hi ha informació rellevant a priori disponible, o quan aquesta informació és subjectiva i desitgem una anàlisi «objectiva», basada exclusivament en les hipòtesis del model i en dades contrastades. Aleshores es fa una *anàlisi de referència* que utilitza conceptes de teoria de la informació per a deduir distribucions finals de referència adients, definides amb l'objectiu d'encapsular inferències sobre les quantitats d'interès, basades únicament en el model suposat i en les dades observades.

En aquest article suposarem que les distribucions de probabilitat es poden descriure a través de llurs funcions de densitat de probabilitat, i no es fa distinció entre una quantitat aleatòria i els valors particulars que pot prendre. Utilitzarem negretes per als vectors aleatòris *observables* (típicament, dades) i lletres gregues en negretes per a vectors aleatòris no observables (típicament, paràmetres); minúscules per a les variables i majúscules per als seus dominis. A més, també utilitzarem la convenció matemàtica estàndard de referir-se a *funcions*, i direm $f(\boldsymbol{x})$ i $g(\boldsymbol{x})$ respectivament, per a f i g de $\boldsymbol{x} \in X$. També, $p(\boldsymbol{\theta} \mid C)$ i $p(\boldsymbol{x} \mid C)$ representaran respectivament *densitats de probabilitat* general dels vectors aleatoris $\boldsymbol{\theta} \in \Theta$ i $\boldsymbol{x} \in X$ sota les condicions C, de manera que $p(\boldsymbol{\theta} \mid C) \geq 0$, $\int_\Theta p(\boldsymbol{\theta} \mid C)\, d\boldsymbol{\theta} = 1$, i $p(\boldsymbol{x} \mid C) \geq 0$, $\int_X p(\boldsymbol{x} \mid C)\, d\boldsymbol{x} = 1$. Aquesta notació, cal reconèixer-ho, no és gaire rigorosa però simplificarà molt l'exposició. Si els vectors aleatoris són discrets, aquestes funcions cal entendre-les com funcions de repartiment de massa, i les integrals cal transformar-les en sumes.

Nom	*Densitat de probabilitat o funció de probabilitat*	*Paràmetre(s)*
Beta	$\mathrm{Be}(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},\ x \in (0,1)$	$\alpha > 0,\ \beta > 0$
Binomial	$\mathrm{Bi}(x \mid n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x},\ x \in \{0, \ldots, n\}$	$n \in \{1, 2, \ldots\},\ \theta \in (0,1)$
Exponencial	$\mathrm{Ex}(x \mid \theta) = \theta\, e^{-\theta x},\ x > 0$	$\theta > 0$
ExpGamma	$\mathrm{Eg}(x \mid \alpha, \beta) = \frac{\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}},\ x > 0$	$\alpha > 0,\ \beta > 0$
Gamma	$\mathrm{Ga}(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},\ x > 0$	$\alpha > 0,\ \beta > 0$
NegBinomial	$\mathrm{Nb}(x \mid r, \theta) = \theta^r \binom{r+x-1}{r-1} (1-\theta)^x,\ x \in \{0, 1, \ldots\}$	$r \in \{1, 2, \ldots\},\ \theta \in (0,1)$
Normal	$\mathrm{N}_k(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{\lvert\Sigma\rvert^{-1/2}}{(2\pi)^{k/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right],\ \boldsymbol{x} \in \Re^k$	$\boldsymbol{\mu} \in \Re^k$, Σ def. pos.
Poisson	$\mathrm{Pn}(x \mid \lambda) = e^{-\lambda} \frac{\lambda^x}{x!},\ x \in \{0, 1, \ldots\}$	$\lambda > 0$
Student	$\mathrm{St}(x \mid \mu, \sigma, \alpha) = \frac{\Gamma(\frac{\alpha+1}{2})}{\Gamma(\frac{\alpha}{2})} \frac{1}{\sigma\sqrt{\alpha\pi}} \left[1 + \frac{1}{\alpha}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-(\alpha+1)/2},\ x \in \Re$	$\mu \in \Re,\ \sigma > 0,\ \alpha > 0$

Taula 1: Notacions per a les densitats de probabilitat i les funcions de probabilitat habituals.

Les funcions de densitat específiques les denotarem pel seu nom. Llavors, si x es una quantitat aleatòria amb distribució normal de mitjana μ i desviació típica σ, denotarem la seva funció de densitat per $\mathrm{N}(x \mid \mu, \sigma^2)$. La taula 1 conté les definicions de les altres distribucions utilitzades en aquest article.

Els mètodes bayesians freqüentment fan ús de la divergència logarítmica, una mesura molt general de la bondat de l'aproximació d'una densitat de probabilitat $p(\boldsymbol{x})$ per a una altra densitat $\hat{p}(\boldsymbol{x})$. La *divergència logarítmica* de la densitat de probabilitat $\hat{p}(\boldsymbol{x})$ del vector aleatori $\boldsymbol{x} \in X$ de la seva densitat de probabilitat autèntica, $p(\boldsymbol{x})$, es defineix per

$$\delta\{\hat{p}(\boldsymbol{x}) \mid p(\boldsymbol{x})\} = \int_X p(\boldsymbol{x}) \log\{p(\boldsymbol{x})/\hat{p}(\boldsymbol{x})\}\, d\boldsymbol{x}.$$

Es pot demostrar que (i) la divergència logarítmica és no negativa (i és zero si, i només si, $\hat{p}(\boldsymbol{x}) = p(\boldsymbol{x})$ quasi per tot arreu), i (ii) $\delta\{\hat{p}(\boldsymbol{x}) \mid p(\boldsymbol{x})\}$ és invariant sota transformacions bijectives de $\boldsymbol{x}$.

Aquest article conté un breu resum dels fonaments matemàtics dels mètodes de l'estadística bayesiana (secció 2), una visió de conjunt del paradigma (secció 3), una descripció de mètodes d'inferència habituals, incloent-hi estimació i test d'hipòtesis (secció 4), una discussió explícita dels mètodes objectius bayesians (secció 5), l'anàlisi detallada d'un cas simplificat (secció 6), i una discussió final que inclou referències per a altres temes no estudiats en la part principal (secció 7).

2 Fonaments

Un element central del paradigma bayesià és l'ús de distribucions de probabilitat per a descriure totes les quantitats rellevants desconegudes, tot interpretant la probabilitat d'un esdeveniment com una mesura condicional de la incertesa, en una escala $[0, 1]$, de l'ocurrència de l'esdeveniment en condicions especificades. Els valors extrems 0 i 1, que són típicament inaccessibles a les aplicacions, descriuen respectivament la impossibilitat i la certesa d'ocurrència de l'esdeveniment. En aquesta interpretació de la probabilitat inclou i estén totes les altres interpretacions de probabilitat. Hi ha dos arguments independents que proven la inevitabilitat matemàtica d'usar distribucions de probabilitat per a descriure incerteses, els quals es resumeixen més endavant en aquesta secció.

2.1 La probabilitat com a mesura de la incertesa condicional

L'estadística bayesiana utilitza la paraula *probabilitat* exactament en el mateix sentit que en el llenguatge ordinari, com a *mesura condicional de la incertesa* associada amb l'ocurrència d'un esdeveniment particular, donades la informació disponible i les hipòtesis acceptades. Així, $\Pr(E \mid C)$ és una mesura de la creença (presumiblement racional) en l'ocurrència de l'*esdeveniment* E sota les *condicions* C. És important remarcar que la probabilitat és *sempre* una

funció de dos arguments: l'esdeveniment E, la incertesa del qual esta essent mesurada, i les condicions C sota les quals es fa el mesurament; les probabilitats «absolutes» no existeixen. En una aplicació típica, hom està interessat en la probabilitat d'algun esdeveniment E donades les *dades* disponibles D, el conjunt d'*hipòtesis* A sobre el mecanisme que ha generat les dades, i el *coneixement* rellevant contextual K que pot estar disponible. Així, $\Pr(E \mid D, A, K)$ ha de ser interpretat com una mesura de creença (presumiblement racional) en l'ocurrència de l'*esdeveniment* E, donades les dades D, les hipòtesis A i qualsevol altre coneixement disponible K, com a mesura de com de «probable» és l'ocurrència d'E en aquestes condicions. Algunes vegades, però no sempre, la probabilitat d'un esdeveniment sota condicions donades pot ser associada amb la freqüència relativa d'esdeveniments «similars» en les mateixes condicions. Els exemples següents tracten d'il·lustrar l'ús de la probabilitat com a mesura condicional de la incertesa.

Diagnòstic probabilístic. Sabem que en una població un 0,2 % de la gent està infectada per un virus determinat. A una persona *escollida a l'atzar* d'aquesta població se li fa un test que dóna positiu al 98 % de la gent infectada i a l'1 % de no-infectats; per tant, si V denota l'esdeveniment que una persona tingui el virus i $+$ denota un resultat positiu, $\Pr(+ \mid V) = 0{,}98$ i $\Pr(+ \mid \overline{V}) = 0{,}01$. Suposem que el resultat del test dóna positiu. Clarament, estem interessats en $\Pr(V \mid +, A, K)$, la *probabilitat* que la persona tingui el virus, atès el resultat positiu, les hipòtesis A sobre el mecanisme probabilístic que genera els resultats del test, i el coneixement disponible K de la prevalença de la infecció en la població sota estudi (descrita aquí per $\Pr(V \mid K) = 0{,}002$). Un exercici elemental de probabilitats amb el teorema de Bayes en la seva forma més simple (vegeu la secció 3) dóna $\Pr(V \mid +, A, K) = 0{,}164$. Noteu que les quatre probabilitats que intervenen en el problema tenen *exactament la mateixa interpretació*: totes són mesures condicionals de la incertesa. A més, $\Pr(V \mid +, A, K)$ és *alhora* una mesura de la incertesa associada amb l'esdeveniment que una persona concreta escollida a l'atzar que ha donat positiva estigui realment infectada, *i* una *estimació* de la proporció de la gent en aquella població (aproximadament 16,4 %) que a la llarga estaran infectats entre tots els que han donat positiu en el test.

Estimació d'una proporció. Es fa un estudi per a estimar la proporció θ d'individus d'una població que tenen una determinada característica. S'analitza una mostra aleatòria de n elements i es troba que r tenen la característica. En general, hom està interessat a utilitzar els resultats de la mostra per a establir regions de $[0, 1]$ on és raonable que es trobi el valor desconegut de θ; aquesta informació s'expressa en termes de *probabilitats* de la forma $\Pr(a < \theta < b \mid r, n, A, K)$, una mesura condicional de la incertesa sobre l'esdeveniment que θ pertanyi a (a, b) *suposada* la informació proporcionada per les dades (r, n), les hipòtesis A sobre el comportament del mecanisme

que ha generat les dades (una mostra aleatòria de n proves de Bernoulli), i qualsevol coneixement rellevant K sobre els valors de θ que hi pugui haver disponible. Per exemple, després d'una enquesta política on 720 ciutadans d'una mostra aleatòria de 1.500 han declarat que estan a favor d'una mesura política, es pot concloure $\Pr(\theta < 0{,}5 \mid 720, 1.500, A, K) = 0{,}933$, que indica una probabilitat de més del 93 % de perdre un referèndum sobre aquest tema. Anàlogament, després d'una recerca mèdica sobre una infecció on es van controlar 100 persones, cap d'elles va resultar que estava infectada es pot afirmar que $\Pr(\theta < 0{,}01 \mid 0, 100, A, K) = 0{,}844$, és a dir, hi ha una probabilitat d'un 84 % que la proporció de gent infectada sigui menor que l'1 %.

Mesura d'una constant física. Un equip de científics que volen establir el valor desconegut d'una constant física μ obtenen dades $D = \{x_1, \ldots, x_n\}$ que són considerades com mesures de μ subjectes a error. Les probabilitats d'interès són, llavors, típicament de la forma $\Pr(a < \mu < b \mid x_1, \ldots, x_n, A, K)$, la *probabilitat* que el valor desconegut de μ (fixat per la natura, però desconegut pels científics) pertanyi a un interval (a, b) atesa la informació proporcionada per les dades D, les hipòtesis A fetes sobre el comportament del mecanisme de mesura i qualsevol coneixement K disponible sobre el valor de la constant μ. Un altre cop, aquestes probabilitats són mesures condicionals de la incertesa que descriuen les conclusions (necessàriament probabilístiques) dels científics sobre el valor veritable de μ, atesa la informació disponible i les hipòtesis acceptades. Per exemple, després d'un experiment a classe per a mesurar el camp gravitatori amb un pèndol (en $\mathrm{m/s^2}$), un estudiant pot afirmar alguna cosa com $\Pr(9{,}788 < g < 9{,}829 \mid D, A, K) = 0{,}95$, que vol dir que, sota el coneixement K i les hipòtesis A, les dades *observades* D indiquen que el veritable valor de g està entre 9,788 i 9,829 amb probabilitat 0,95, una mesura de la incertesa condicional en una escala $[0, 1]$. Naturalment, això és compatible amb el fet que el valor del camp gravitatori al laboratori pot ser conegut amb molta precisió a partir de la bibliografia o d'experiments anteriors molt exactes, però es pot haver demanat a l'estudiant de *no* utilitzar aquesta informació com a part del coneixement acceptat K. Sota algunes condicions, també és cert que si el mateix *procediment* fos utilitzat per molts altres estudiants i obtinguessin dades semblants, llurs intervals cobririen el valor veritable de g aproximadament el 95 % dels casos, i proporcionarien així una forma de *calibració* per a l'afirmació probabilística de l'estudiant (vegeu la secció 5.2).

Predicció. Es fa un experiment per a comptar el nombre r de vegades que ocorre un esdeveniment E en n replicacions d'una situació ben definida; s'observa que E ocorre r_i vegades de la replicació i, i es desitja predir el nombre de vegades r que E ocorrerà en una situació similar futura. Aquest és un problema de *predicció* sobre el valor d'una quantitat *observable* (discreta) r, donada la informació proporcionada per les dades D, les hipòtesis acceptades A sobre el mecanisme probabilístic que han generat els r_i, i qualsevol conei-

xement disponible K. Per tant, només es demana el càlcul de les probabilitats $\{\Pr(r \mid r_1, \ldots, r_n, A, K)\}$, per a $r = 0, 1, \ldots$ Per exemple, l'enginyer de qualitat d'una empresa que fabrica sistemes de seguretat de cotxes pot afirmar una cosa de l'estil $\Pr(r = 0 \mid r_1 = \cdots = r_{10} = 0, A, K) = 0{,}953$, després d'observar la producció completa d'*airbags* i que no s'han esdevingut queixes dels clients durant $n = 10$ mesos consecutius. Aquest nombre s'ha de mirar com una mesura, en una escala $[0, 1]$, de la incertesa condicional, ateses les dades observades, les hipòtesis acceptades i el coneixement contextual, associats amb l'esdeveniment que no hi haurà cap queixa en la producció dels *airbags* del proper mes i, si les condicions es mantenen constants, aquesta és també una estimació de la proporció al llarg del mesos.

És natural proposar un problema similar amb observables de tipus contínu. Per exemple, després de mesurar una magnitud contínua en cadascun dels n elements escollits a l'atzar en una població, es desitja predir la proporció d'ítems de la població completa que tenen determinades especificacions. Per exemple, després de mesurar la resistència al trencament de $\{x_1, \ldots, x_{10}\}$ de 10 cinturons de seguretat escollits a l'atzar per comprovar si compleixen l'especificació de resistir més de 26 kN, l'enginyer de qualitat pot informar que $\Pr(x > 26 \mid x_1, \ldots, x_{10}, A, K) = 0{,}9987$. Això s'ha d'interpretar com una mesura, en una escala $[0, 1]$, de la incertesa condicional (donades les dades observades, les hipòtesis acceptades i el coneixement contextual) associada amb l'esdeveniment que un cinturó de seguretat escollit a l'atzar suporti almenys 26 kN. Si les condicions de producció es mantenen constants, pot ser també una estimació de la proporció de cinturons de seguretat que compleixen l'especificació demanada.

La informació addicional sobre observacions futures la proporcionen els covariants relacionats. Per exemple, després d'observar els resultats $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ que corresponen a una successió $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ de condicions de producció diferents, volem predir el resultat $\boldsymbol{y}$ que correspondria a un conjunt particular $\boldsymbol{x}$ de condicions de producció. Per exemple, es vol que els valors de la viscositat de la llet condensada comercial estiguin entre uns valors especificats, a i b; després de mesurar les viscositats $\{y_1, \ldots, y_n\}$ que corresponen a mostres de llet condensada produïda sota condicions físiques diferents, $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, els enginyers de producció necessitaran probabilitats del tipus $\Pr(a < y < b \mid \boldsymbol{x}, (y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n), A, K)$. Es tracta d'una mesura condicional de la incertesa (suposades sempre unes dades observades, hipòtesis acceptades i coneixement contextual) associades amb l'esdeveniment que la llet condensada produïda sota condicions $\boldsymbol{x}$ complirà, de fet, les condicions de viscositat especificades.

2.2 Inferència estadística i teoria de la decisió

La teoria de la decisió no proporciona només una metodologia precisa per a tractar amb problemes de decisió sota incertesa, sinó que amb la seva sòlida base axiomàtica també proporciona un poderós argument per a la força lògica de l'enfocament bayesià. A continuació resumim l'argument fonamental.

Tenim un problema de decisió quan hi ha dues o més accions possibles; designem $\mathcal{A}$ com la col·lecció de les accions possibles. A més, per a cada $a \in \mathcal{A}$, sigui Θ_a el conjunt dels *esdeveniments rellevants* que poden afectar el resultat d'escollir a, i sigui $c(a, \boldsymbol{\theta}) \in C_a$, $\boldsymbol{\theta} \in \Theta_a$, la *conseqüència* d'haver escollit l'acció a quan ocorre l'esdeveniment $\boldsymbol{\theta}$. La col·lecció de parelles $\{(\Theta_a, C_a), a \in \mathcal{A}\}$ descriu l'*estructura* del problema de decisió. Sense pèrdua de generalitat, podem suposar que les possibles accions són mútuament excloents, ja que en cas contrari, hom pot treballar en el producte cartesià adient.

S'han proposat diferents conjunts de principis per a concretar una mínima col·lecció de regles lògiques que puguin ser raonablement demanades per a una presa de decisió «racional». Totes són axiomes amb un fort contingut intuïtiu; com exemples esmentem la *transitivitat* de preferències (si $a_1 > a_2$ donat C, i $a_2 > a_3$ donat C, llavors $a_1 > a_3$ donat C), i el *principi de seguretat* (si $a_1 > a_2$ donat C i E, i $a_1 > a_2$ donat C i $\overline{E}$, llavors $a_1 > a_2$ donat C). Notem que aquestes regles no pretenen ser una descripció de la presa de decisions humanes sinó un conjunt *normatiu* de principis que hauria de seguir qualsevol persona que aspira a ser un decisor coherent.

Naturalment hi ha diferents opcions per al conjunt de principis acceptables, però totes condueixen bàsicament a la mateixa conclusió:

a) Les preferències entre conseqüències han de ser mesurades amb una funció d'*utilitat* a valors reals, fitada $U(c) = U(a, \boldsymbol{\theta})$, que especifiqui en alguna escala numèrica la seva desitjabilitat.

b) La incertesa d'esdeveniments rellevants ha de ser mesurada amb un conjunt de distribucions de *probabilitat* $\{(p(\boldsymbol{\theta} \mid C, a), \boldsymbol{\theta} \in \Theta_a), a \in \mathcal{A}\}$ que descriuen la seva plausibilitat donades les condicions C sota les quals la decisió s'ha de prendre.

c) La desitjabilitat de les accions disponibles es mesura amb llur corresponent *utilitat esperada*

$$\overline{U}(a \mid C) = \int_{\Theta_a} U(a, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid C, a)\, d\boldsymbol{\theta}, \quad a \in \mathcal{A}\,. \tag{1}$$

Sovint és convenient treballar en termes de la funció de *pèrdua* no negativa definida per

$$L(a, \boldsymbol{\theta}) = \sup_{a \in \mathcal{A}} \{U(a, \boldsymbol{\theta})\} - U(a, \boldsymbol{\theta})\,, \tag{2}$$

que mesura directament, com a funció de $\boldsymbol{\theta}$, el «càstig» per escollir una acció incorrecta. Es mesura la no-desitjabilitat relativa d'una acció disponible $a \in \mathcal{A}$ per la seva *pèrdua esperada*

$$\overline{L}(a \mid C) = \int_{\Theta_a} L(a, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid C, a)\, d\boldsymbol{\theta}, \quad a \in \mathcal{A}\,. \tag{3}$$

Noteu que, en particular, l'argument anterior estableix la necessitat de quantificar la incertesa sobre totes les quantitats rellevants desconegudes (els actuals valors de les $\boldsymbol{\theta}$), i especifica que aquesta quantificació *cal* que tingui

l'estructura matemàtica de les distribucions de probabilitat. És a dir, en termes de probabilitats condicionades sota les circumstàncies C en les quals la decisió ha de ser presa, que típicament, però no necessàriament, inclou els resultats D d'alguna dada rellevant experimental o observada.

S'ha argumentat que el desenvolupament descrit anteriorment (que no es qüestiona quan cal prendre decisions), no s'aplica a problemes d'inferència estadística quan no es preveuen decisions a prendre. Ara bé, hi ha dos poderosos contraarguments, que són: *a*) es considera interessant analitzar un problema d'inferència estadística perquè *pot* ajudar a prendre decisions raonables (com Ramsey explicà els anys trenta, un grumoll d'arsènic és verinós perquè *pot* matar algú, no perquè hagi matat algú), i *b*) s'ha demostrat (ho va fer Bernardo els anys setanta) que la inferència estadística sobre $\boldsymbol{\theta}$ *té*, de fet, l'estructura matemàtica d'un problema de decisió, on la classe d'alternatives és l'espai funcional

$$\mathcal{A} = \left\{ p(\boldsymbol{\theta} \mid D); \quad p(\boldsymbol{\theta} \mid D) > 0, \int_{\Theta} p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta} = 1 \right\} \tag{4}$$

de distribucions de probabilitat condicionades de $\boldsymbol{\theta}$ ateses les dades, i la funció d'utilitat és una mesura de la quantitat d'informació sobre $\boldsymbol{\theta}$ que es pot esperar que les dades proporcionaran.

2.3 Intercanviabilitat i teorema de representació

Les dades disponibles sovint prenen la forma d'un conjunt $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ d'observacions «homogènies», en el sentit exacte que només els seus *valors* importen i no l'*ordre* en què apareixen. Formalment, aquesta idea és captada amb la noció d'*intercanviabilitat*. D'un conjunt de vectors aleatoris $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ es diu que és intercanviable si la seva distribució conjunta és invariant sota permutacions. Una successió infinita $\{\boldsymbol{x}_j\}$ de vectors aleatoris és intercanviable si totes les seves subsuccessions finites són intercanviables. Noteu que, en particular, qualsevol mostra aleatòria a partir de qualsevol model és intercanviable. El concepte d'intercanviabilitat, introduït per de Finetti els anys trenta, és central en el pensament estadístic modern. De fet, el *teorema de representació* general implica que si s'assumeix que un conjunt d'observacions és un subconjunt d'una successió intercanviable, aleshores constitueix *una mostra aleatòria* d'algun model de probabilitat $\{p(\boldsymbol{x} \mid \boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, $\boldsymbol{x} \in X$, retolat per un *vector de paràmetres* $\boldsymbol{\omega}$; a més, aquest paràmetre $\boldsymbol{\omega}$ està *definit* com el límit (quan $n \to \infty$) d'una funció de les observacions. La informació disponible sobre el valor de $\boldsymbol{\omega}$ en les condicions imperants C és *necessàriament* descrita per *alguna* distribució de probabilitat $p(\boldsymbol{\omega} \mid C)$.

Per exemple, en el cas d'una successió $\{x_1, x_2, \ldots\}$ de quantitats aleatòries intercanviables dicotòmiques $x_j \in \{0, 1\}$, el teorema de representació demostrat per de Finetti estableix que la distribució conjunta de $(x_1, \ldots, x_n)$ té una

representació integral de la forma

$$p(x_1,\ldots,x_n \mid C) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}\, p(\theta \mid C)\, d\theta, \qquad \theta = \lim_{n\to\infty} \frac{r}{n}\ , \tag{5}$$

on $r = \sum x_j$ és el nombre de proves positives. Aquesta és precisament la distribució conjunta d'una família de proves de Bernoulli (condicionalment) independents amb paràmetre θ, sobre el qual es demostra que existeix alguna distribució de probabilitat $p(\theta \mid C)$. Més generalment, per a successions de quantitats aleatòries arbitràries $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\}$, la intercanviabilitat porta a representacions integrals de la forma

$$p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n \mid C) = \int_{\boldsymbol{\Omega}} \prod_{i=1}^{n} p(\boldsymbol{x}_i \mid \boldsymbol{\omega})\, p(\boldsymbol{\omega} \mid C)\, d\boldsymbol{\omega}\,, \tag{6}$$

on $\{p(\boldsymbol{x} \mid \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$ denota algun *model* probabilístic, $\boldsymbol{\omega}$ és el límit quan $n \to \infty$ d'alguna funció $f_n(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ de les observacions, i $p(\boldsymbol{\omega} \mid C)$ és alguna distribució de probabilitat sobre $\boldsymbol{\Omega}$. Aquesta formulació inclou el modelat «no paramètric» (distribució lliure), on $\boldsymbol{\omega}$ pot indexar, per exemple, totes les distribucions de probabilitat contínues sobre X. Noteu que $p(\boldsymbol{\omega} \mid C)$ *no* descriu una possible variabilitat de $\boldsymbol{\omega}$ (ja que $\boldsymbol{\omega}$ normalment serà un vector fix però *desconegut*), sinó que descriu la incertesa associada amb el seu valor actual.

Sota condicions adients, la intercanviabilitat és una hipòtesi molt general, una extensió potent del concepte tradicional de *mostra aleatòria*. De fet, moltes anàlisis estadístiques assumeixen directament que les dades (o subconjunts de les dades) són una mostra aleatòria d'observacions condicionalment independents d'algun model de probabilitat, per tant, $p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n \mid \boldsymbol{\omega}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i \mid \boldsymbol{\omega})$; però *qualsevol* mostra aleatòria és intercanviable, atès que $\prod_{i=1}^{n} p(\boldsymbol{x}_i \mid \boldsymbol{\omega})$ és òbviament invariant sota permutacions. Noteu que les observacions en una mostra aleatòria només són independents *condicionalment* al valor del paràmetre $\boldsymbol{\omega}$; com elegantment diu Lindley, el *mantra* que les observacions $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ en una mostra aleatòria són independents és ridícul quan s'utilitza per a deduir $\boldsymbol{x}_{n+1}$. Noteu també que, sota la intercanviabilitat, el teorema general de representació proporciona un *teorema d'existència* d'una distribució de probabilitat $p(\boldsymbol{\omega} \mid C)$ a l'espai de paràmetres Ω, i que aquest és un argument que només reposa sobre la teoria matemàtica de la probabilitat.

Una altra conseqüència important de la intercanviabilitat és que proporciona una *definició* formal del paràmetre $\boldsymbol{\omega}$ que retola el model com el límit, quan $n \to \infty$, d'*alguna* funció $f_n(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ de les observacions; la funció f_n òbviament depèn del model proposat i de la parametrització escollida. Per exemple, en el cas d'una successió de proves de Bernoulli, el paràmetre θ està *definit* com el límit, quan $n \to \infty$, de les freqüències relatives r/n. Per tant, sota intercanviabilitat, la frase «el veritable valor de $\boldsymbol{\omega}$» té un sentit ben definit, encara que només és asimptòticament comprovable. A més,

si dos models diferents tenen paràmetres que per definició estan funcionalment relacionats, aleshores les distribucions finals corresponents poden ser comparades amb sentit, perque es refereixen a quantitats funcionalment relacionades. Per exemple, si suposem que un subconjunt finit $\{x_1, \ldots, x_n\}$ d'una successió intercanviable d'observacions enteres és una mostra aleatòria d'una distribució de Poisson $\text{Po}(x \mid \lambda)$, amb $\text{E}[x \mid \lambda] = \lambda$, aleshores λ és *defineix* com $\lim_{n\to\infty}\{\overline{x}_n\}$, on $\overline{x}_n = \sum_j x_j/n$; de manera similar, si per a algun enter no nul fixat r, les mateixes dades se suposa que són una mostra aleatòria d'una binomial negativa $\text{Nb}(x \mid r, \theta)$, amb $\text{E}[x \mid \theta, r] = r(1-\theta)/\theta$, llavors θ es *defineix* com $\lim_{n\to\infty}\{r/(\overline{x}_n + r)\}$. D'on es dedueix $\theta \equiv r/(\lambda + r)$ i, per tant, θ i $r/(\lambda + r)$ pot ser tractat com la *mateixa* quantitat (desconeguda), sigui quina sigui, com, per exemple, quan comparem els mèrits relatius de dos models de probabilitat alternatius.

3 El paradigma bayesià

L'anàlisi estadística d'unes dades observades D normalment comença amb una avaluació informal *descriptiva*, que s'utilitza per a suggerir un *model probabilístic* formal $\{p(D \mid \boldsymbol{\omega}),\ \boldsymbol{\omega} \in \Omega\}$ que suposem que representa, per a algun valor (desconegut) de $\boldsymbol{\omega}$, el mecanisme probabilístic que ha generat les dades observades D. Els arguments indicats a la secció 2 estableixen la necessitat lògica d'assignar una distribució de probabilitat *inicial* $p(\boldsymbol{\omega} \mid K)$ sobre l'espai de paràmetres Ω, que descrigui el coneixement disponible K sobre el valor de $\boldsymbol{\omega}$ anterior al fet que les dades fossin observades. Aleshores, a partir de càlculs estàndards de la teoria de probabilitat, si el model probabilístic és correcte, tota la informació disponible sobre el valor de $\boldsymbol{\omega}$ després que les dades D han estat observades, està contingut en la corresponent distribució *final*, la densitat de probabilitat de la qual, $p(\boldsymbol{\omega} \mid D, A, K)$, s'obté de manera immediata a partir del teorema de Bayes,

$$p(\boldsymbol{\omega} \mid D, A, K) = \frac{p(D \mid \boldsymbol{\omega})\, p(\boldsymbol{\omega} \mid K)}{\int_\Omega p(D \mid \boldsymbol{\omega})\, p(\boldsymbol{\omega} \mid K)\, d\boldsymbol{\omega}}\ , \tag{7}$$

on A representa les hipòtesis fetes sobre el model de probabilitat. És aquest ús sistemàtic del teorema de Bayes per a incorporar la informació proporcionada per les dades que justifica l'adjectiu *bayesià* pel qual es coneix normalment el paradigma. A partir del teorema de Bayes és evident que qualsevol valor de $\boldsymbol{\omega}$ amb densitat inicial zero tindrà densitat final zero. Així, normalment se suposa (mitjançant la restricció adient, si cal, de l'*espai de paràmetres* Ω) que les distribucions inicials són *estrictament positives* (tal com Savage deia, manté la ment oberta o, almenys, entreoberta). Per simplificar la presentació, suprimirem la referència a les hipòtesis acceptades A i el coneixement disponible K, però cal sempre tenir present que *totes* les afirmacions sobre $\boldsymbol{\omega}$ suposat D són *també* condicionals a A i K.

1 EXEMPLE *(Inferència bayesiana amb espai de paràmetres finit.)* Sigui $p(D \mid \theta)$, $\theta \in \{\theta_1, \ldots, \theta_m\}$, el mecanisme probabilístic que suposarem que ha generat les dades observades D, de manera que θ només pot prendre un nombre *finit* de valors. A partir de la forma finita del teorema de Bayes, i ometent en la notació les condicions imposades, la probabilitat final de θ_i després que les dades D han estat observades és

$$\Pr(\theta_i \mid D) = \frac{p(D \mid \theta_i)\ \Pr(\theta_i)}{\sum_{j=1}^{m} p(D \mid \theta_j)\ \Pr(\theta_j)}\,, \quad i = 1, \ldots, m\,. \tag{8}$$

Per a qualsevol distribució inicial $p(\theta) = \{\Pr(\theta_1), \ldots, \Pr(\theta_m)\}$ que descrigui el coneixement disponible sobre el valor de θ, $\Pr(\theta_i \mid D)$ mesura com de probable ha de ser jutjat θ_i, suposats el coneixement inicial descrit per la distribució inicial i la informació proporcionada per les dades D.

Una aplicació important i freqüent d'aquesta tècnica tan simple ens la proporciona el diagnòstic probabilístic. Per exemple, considerem la situació on sabem, a partir de la recerca del laboratori, que un test dissenyat per a detectar un virus dóna un resultat positiu en el 98% de la gent infectada i en l'1% de no infectats. Llavors, la probabilitat final que una persona que ha donat positiu en el test estigui infectada ve donada per $\Pr(V \mid +) = (0{,}98\,p)/\{0{,}98\,p + 0{,}01\,(1-p)\}$ com a funció de $p = \Pr(V)$, (probabilitat inicial d'estar una persona infectada: *prevalença* de la infecció a la població sota estudi). La figura 1 mostra $\Pr(V \mid +)$ com a funció de $\Pr(V)$.

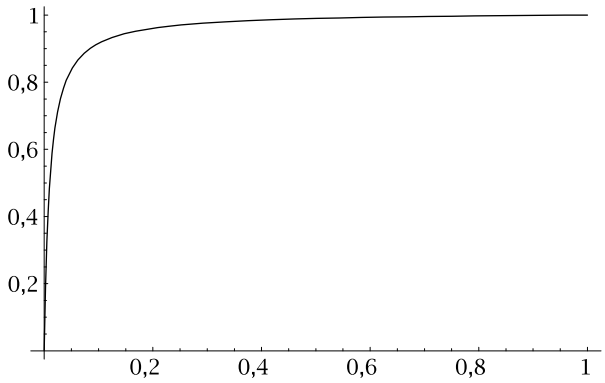

FIGURA 1: Probabilitat final d'infecció $\Pr(V \mid +)$ atès un test positiu, com a funció de la probabilitat inicial d'infecció $\Pr(V)$.

Com hom pot esperar, la probabilitat final és zero només si la probabilitat inicial ho és (és a dir, quan és *conegut* que ningú de la població no té la infecció) i és 1 només si la probabilitat inicial ho és (és a dir, quan és *conegut* que la població està universalment infectada). Noteu que si la infecció és

poc freqüent, aleshores la probabilitat final que una persona escollida a l'atzar estigui infectada serà relativament baixa, fins i tot si el test és positiu. De fet, per exemple, per a $\Pr(V) = 0{,}002$, es troba $\Pr(V \mid +) = 0{,}164$, de manera que en una població on només el 0,2 % dels individus estan infectats, només el 16,4 % de la gent a qui el test ha donat positiu en una mostra aleatòria estarà realment infectada: la majoria de positius seran *falsos* positius.

En aquesta secció descriurem amb algun detall el procés d'aprenentatge descrit pel teorema de Bayes, en discutirem la implementació en presència de paràmetres marginals, mostrarem com es pot utilitzar per a predir el valor d'observacions futures i analitzarem el comportament en una mostra gran.

3.1 El procés d'aprenentatge

En el paradigma bayesià, el procés d'aprenentatge a partir de les dades és sistemàticament implementat fent ús del teorema de Bayes per a combinar la informació inicial disponible amb la informació proporcionada per les dades per a produir la distribució final. Els càlculs de densitats finals es faciliten observant que el teorema de Bayes pot ser simplement expressat com

$$p(\boldsymbol{\omega} \mid D) \propto p(D \mid \boldsymbol{\omega})\, p(\boldsymbol{\omega})\,, \tag{9}$$

(on $\propto$ vol dir *proporcional a* i on, per simplicitat, la hipòtesi acceptada A i el coneixement disponible K s'ometen de la notació), i la constant de proporcionalitat que falta, $[\int_\Omega p(D \mid \boldsymbol{\omega})\, p(\boldsymbol{\omega})\, d\boldsymbol{\omega}]^{-1}$, pot ser deduïda del fet que $p(\boldsymbol{\omega} \mid D)$, una densitat de probabilitat, ha de tenir integral 1. D'aquí, per a identificar la forma d'una distribució final només cal identificar un *nucli* de la corresponent densitat de probabilitat, que és una funció $k(\boldsymbol{\omega}|D)$ de manera que $p(\boldsymbol{\omega} \mid D) = c(D)\, k(\boldsymbol{\omega}|D)$ per a algun $c(D)$ que no inclou $\boldsymbol{\omega}$. En els exemples que segueixen s'utilitzarà sistemàticament aquesta tècnica.

Una *funció inicial impròpia* es defineix com una funció positiva $\pi(\boldsymbol{\omega})$ tal que $\int_\Omega \pi(\boldsymbol{\omega})\, d\boldsymbol{\omega}$ és no finita. L'equació (9), l'expressió formal del teorema de Bayes, continua essent tècnicament correcta si $p(\boldsymbol{\omega})$ és reemplaçat per una funció inicial impròpia $\pi(\boldsymbol{\omega})$ suposant que la constant de proporcionalitat existeix, i dóna així una densitat final *pròpia* ben definida $\pi(\boldsymbol{\omega} \mid D) \propto p(D \mid \boldsymbol{\omega})\pi(\boldsymbol{\omega})$. Més endavant provarem (secció 5) que el teorema de Bayes també continua essent filosòficament correcte si $p(\boldsymbol{\omega})$ és reemplaçat per una funció inicial de referència «no informativa» (normalment impròpia) $\pi(\boldsymbol{\omega})$ escollida de manera adient.

Considerada com a funció de $\boldsymbol{\omega}$, $l(\boldsymbol{\omega}, D) = p(D \mid \boldsymbol{\omega})$ s'anomena normalment *funció de versemblança*. Així, el teorema de Bayes es pot expressar simplement en paraules dient que *la distribució final és proporcional a la versemblança multiplicada per la distribució inicial*. Es dedueix de l'equació (9) que, utilitzant la *mateixa* distribució inicial $p(\boldsymbol{\omega})$, dos conjunts de dades diferents D_1 i D_2, amb models probabilístics $p_1(D_1 \mid \boldsymbol{\omega})$ i $p_2(D_2 \mid \boldsymbol{\omega})$ possiblement diferents però que donin funcions de versemblança *proporcionals*, produiran

idèntiques distribucions finals per a $\boldsymbol{\omega}$. Aquesta conseqüència immediata del teorema de Bayes ha estat proposada com un principi independent, el *principi de versemblança*, i molta gent el considera com un requeriment evident per a una inferència estadística raonable. En particular, per a una distribució inicial donada $p(\boldsymbol{\omega})$, la distribució final no depèn del conjunt dels possibles valors de les dades o *espai de resultats*. Noteu, però, que el principi de versemblança només s'aplica a inferències sobre el vector de paràmetres $\boldsymbol{\omega}$ una vegada que les dades han estat obtingudes. La consideració de l'espai de resultats és essencial, per exemple, en la crítica de models, en el disseny d'experiments, en la derivació de distribucions predictives o (vegeu la secció 5) en la construcció de procediments bayesians objectius.

Naturalment, els termes *distribucions inicial i final* són només *relatius* a un conjunt particular de dades. Tal com hom espera de la coherència induïda per la teoria de la probabilitat, si les dades $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ són presentades seqüencialment, el resultat final serà el mateix tant si les dades són processades globalment com seqüèncialment. De fet, $p(\boldsymbol{\omega} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i+1}) \propto p(\boldsymbol{x}_{i+1} \mid \boldsymbol{\omega})\, p(\boldsymbol{\omega} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_i)$, per a $i = 1, \ldots, n-1$, i, per tant, la «distribució final» en una etapa donada esdevé la «distribució inicial» a la següent.

En moltes situacions, la distribució final és «més estreta» que la inicial ja que, en la majoria de casos, $p(\boldsymbol{\omega} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i+1})$ estarà més concentrada al voltant del valor veritable de $\boldsymbol{\omega}$ que $p(\boldsymbol{\omega} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_i)$. Però aquest no és sempre el cas: de tant en tant, una observació «sorprenent» augmentarà, en lloc de disminuir, la incertesa sobre el valor de $\boldsymbol{\omega}$. Per exemple, en diagnòstic probabilístic, una distribució de probabilitat final puntual (sobre les possibles causes $\{\omega_1, \ldots, \omega_k\}$ d'una síndrome) que descrigui un «clar» diagnòstic de malaltia ω_i (que és, una probabilitat final gran per a ω_i) normalment s'actualitzarà en una distribució de probabilitat final menys concentrada sobre $\{\omega_1, \ldots, \omega_k\}$ si una nova anàlisi clínica proporciona dades que eren improbables sota ω_i.

Per a un model de probabilitat donat, es pot trobar que una funció particular de les dades $\boldsymbol{t} = \boldsymbol{t}(D)$ és un estadístic suficient en el sentit que donat el model $\boldsymbol{t}(D)$ conté tota la informació sobre $\boldsymbol{\omega}$ que està disponible en D. Formalment, $\boldsymbol{t} = \boldsymbol{t}(D)$ és suficient si (i només si) existeixen funcions no negatives f i g de manera que la funció de versemblança pot ser factoritzada en la forma $p(D \mid \boldsymbol{\omega}) = f(\boldsymbol{\omega}, \boldsymbol{t})g(D)$. Un estadístic suficient sempre existeix, ja que $\boldsymbol{t}(D) = D$ és òbviament suficient; però un estadístic suficient molt més simple, amb una dimensió fixada independent de la grandària mostral, també existeix sovint. De fet, aquest és el cas quan el model de probabilitat pertany a la *família exponencial generalitzada*, que inclou molts dels models probabilístics més freqüentment usats. Es demostra fàcilment que, si $\boldsymbol{t}$ és suficient, la distribució final de $\boldsymbol{\omega}$ només depèn de la dades D a través de $\boldsymbol{t}(D)$, i pot ser directament calculat en termes de $p(\boldsymbol{t} \mid \boldsymbol{\omega})$, i, per tant, $p(\boldsymbol{\omega} \mid D) = p(\boldsymbol{\omega} \mid \boldsymbol{t}) \propto p(\boldsymbol{t} \mid \boldsymbol{\omega})\, p(\boldsymbol{\omega})$.

Naturalment, per a unes dades i hipòtesis fixades sobre el model, diferents distribucions inicials condueixen a diferents distribucions finals. De fet, el teorema de Bayes pot ser descrit com una màquina de transformar probabilitats dirigida per les dades, que a partir de distribucions inicials (que descriuen el

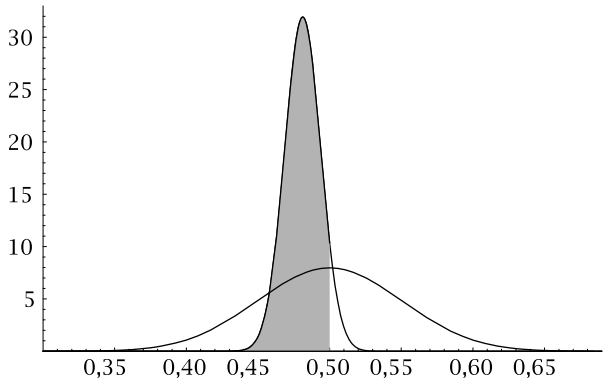

FIGURA 2: Densitats inicial i final de la proporció θ de ciutadans que votaran a favor en el referèndum.

coneixement inicial) produeix distribucions finals (que representen la combinació de coneixement inicial i dades). És important analitzar quan canvis en la distribució inicial indueixen canvis importants en la distribució final. Les distribucions finals basades en funcions inicials de referència «no informatives» tenen un paper central en aquest context d'*anàlisi de sensibilitat*. La investigació de la sensibilitat de la distribució final a canvis en la distribució inicial és un ingredient important d'una anàlisi de la sensibilitat comprensiu dels resultats finals de *totes* les hipòtesis acceptades que qualsevol estudi estadístic responsable ha d'incloure.

2 EXEMPLE *(Inferència sobre un paràmetre binomial.)* Si les dades D consisteixen en n observacions Bernoulli amb paràmetre θ que contenen r proves positives, llavors $p(D \mid \theta, n) = \theta^r(1-\theta)^{n-r}$, tals que $\boldsymbol{t}(D) = \{r, n\}$ és suficient. Suposem que el coneixement inicial sobre θ és descrita per una distribució Beta $\mathrm{Be}(\theta \mid \alpha, \beta)$, de manera que $p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$. A partir del teorema de Bayes, la densitat final de θ és

$$p(\theta \mid r, n, \alpha, \beta) \propto \theta^r(1-\theta)^{n-r}\,\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{r+\alpha-1}(1-\theta)^{n-r+\beta-1},$$

la distribució Beta $\mathrm{Be}(\theta \mid r+\alpha, n-r+\beta)$.

Suposem, per exemple, que en vista d'enquestes anteriors, la informació disponible sobre la proporció θ de ciutadans que votaran per una determinada mesura política en un referèndum és descrita per una distribució Beta $\mathrm{Be}(\theta \mid 50, 50)$, és a dir, és igual de probable que el referèndum es guanyi com que es perdi, i que s'estima que la probabilitat que qualsevol banda guanyi menys del 60 % dels vots és 0,95.

Es pren una mostra aleatòria de grandària 1.500, on només 720 ciutadans declaren estar en favor de la mesura proposada. Utilitzant els resultats ante-

riors, la distribució final corresponent és $\text{Be}(\theta \mid 730, 790)$. Aquestes densitats inicial i final estan dibuixades a la figura 2; pot observar-se que, tal com podia esperar-se, l'efecte de les dades és reduir dràsticament la incertesa inicial sobre el valor de θ i, d'aquí, sobre el resultat del referèndum. Més precisament, $\Pr(\theta < 0{,}5 \mid 720, 1.500, H) = 0{,}933$ (regió ombrejada a la figura 2) per tant, després d'incloure la informació de la mostra, la probabilitat que el referèndum es perdi ha de ser estimada en el 93 %.

A continuació estudiarem la situació general on el vector d'interès no és tot el vector de paràmetres $\boldsymbol{\omega}$, sinó una funció $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega})$ de dimensió possiblement inferior a $\boldsymbol{\omega}$. Siguin D les dades observades i $\{p(D \mid \boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ un model de probabilitat que suposarem que descriu el mecanisme probabilístic que ha generat D, i sigui $p(\boldsymbol{\omega})$ una distribució de probabilitat que descriu tota la informació disponible sobre el valor de $\boldsymbol{\omega}$, i volem fer inferències sobre el valor d'una funció del paràmetre original $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \Theta$ a partir de les dades D. Qualsevol conclusió sobre el valor del *vector d'interès* $\boldsymbol{\theta}$ estarà continguda en la distribució de probabilitat final $p(\boldsymbol{\theta} \mid D)$, que és condicional a les dades observades D i, naturalment també depèn, encara que no es mostri explícitament en la notació, del model $\{p(D \mid \boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, i de la informació inicial disponible encapsulada per $p(\boldsymbol{\omega})$. La distribució final demanada $p(\boldsymbol{\theta} \mid D)$ es troba amb una utilització estàndard del càlcul de probabilitats. De fet, amb el teorema de Bayes, $p(\boldsymbol{\omega} \mid D) \propto p(D \mid \boldsymbol{\omega})\, p(\boldsymbol{\omega})$. A més, sigui $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\omega}) \in \Lambda$ alguna altra funció dels paràmetres originals de manera que $\boldsymbol{\psi} = \{\boldsymbol{\theta}, \boldsymbol{\lambda}\}$ és una transformació bijectiva de $\boldsymbol{\omega}$, i sigui $J(\boldsymbol{\omega}) = (\partial \boldsymbol{\psi} / \partial \boldsymbol{\omega})$ la corresponent matriu jacobiana. Naturalment, la introducció de $\boldsymbol{\lambda}$ no és necessària si $\boldsymbol{\theta}(\boldsymbol{\omega})$ és una transformació bijectiva de $\boldsymbol{\omega}$. Mitjançant tècniques estàndard de canvi de variables, la densitat final de $\boldsymbol{\psi}$ és

$$p(\boldsymbol{\psi} \mid D) = p(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid D) = \left[\frac{p(\boldsymbol{\omega} \mid D)}{\mid J(\boldsymbol{\omega}) \mid} \right]_{\boldsymbol{\omega} = \boldsymbol{\omega}(\boldsymbol{\psi})}, \tag{10}$$

i la densitat final de $\boldsymbol{\theta}$ és la densitat *marginal* corresponent, obtinguda per integració sobre el *paràmetre marginal* $\boldsymbol{\lambda}$

$$p(\boldsymbol{\theta} \mid D) = \int_{\Lambda} p(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid D)\, d\boldsymbol{\lambda}\,. \tag{11}$$

Noteu que l'eliminació de paràmetres marginals no desitjats és una simple integració en el paradigma bayesià, però un problema difícil (sovint polèmic) per a l'estadística convencional.

A vegades, es restringeix molt el rang de possibles valors de $\boldsymbol{\omega}$ per consideracions contextuals. Si és conegut que $\boldsymbol{\omega}$ pertany a $\Omega_c \subset \Omega$, la distribució inicial és positiva només en Ω_c i, utilitzant el teorema de Bayes, és immediat que la distribució final restringida és

$$p(\boldsymbol{\omega} \mid D, \boldsymbol{\omega} \in \Omega_c) = \frac{p(\boldsymbol{\omega} \mid D)}{\int_{\Omega_c} p(\boldsymbol{\omega} \mid D)}\,, \quad \boldsymbol{\omega} \in \Omega_c\,, \tag{12}$$

i òbviament s'anul·la si $\boldsymbol{\omega} \notin \Omega_c$. Així, per a incorporar una restricció sobre els valors possibles dels paràmetres, només cal *renormalitzar* la distribució final no restringida al conjunt de valors $\Omega_c \subset \Omega$ del paràmetre que compleixen la condició demanada. La incorporació de restriccions conegudes sobre els valors del paràmetre, una simple renormalització en el paradigma bayesià, és un altre problema difícil per a l'estadística convencional.

3 EXEMPLE *(Inferència sobre paràmetres normals.)* Sigui $D = \{x_1, \ldots, x_n\}$ una mostra aleatòria d'una distribució normal $N(x \mid \mu, \sigma^2)$. La funció de versemblança corresponent és proporcional a $\sigma^{-n} \exp[-n\{s^2 + (\overline{x} - \mu)^2\}/(2\sigma^2)]$, amb $n\overline{x} = \sum_i x_i$, i $ns^2 = \sum_i (x_i - \overline{x})^2$. Es pot demostrar (vegeu la secció 5) que l'absència d'informació inicial sobre els valors de μ i σ pot ser descrita formalment amb una funció conjunta inicial uniforme en ambdues μ i $\log(\sigma)$, que és la funció inicial (impròpia) $p(\mu, \sigma) = \sigma^{-1}$. D'acord amb el teorema de Bayes, la final conjunta corresponent és

$$p(\mu, \sigma \mid D) \propto \sigma^{-(n+1)} \exp[-n\{s^2 + (\overline{x} - \mu)^2\}/(2\sigma^2)]\,. \tag{13}$$

Així, a partir d'una integral Gamma en termes de $\lambda = \sigma^{-2}$ integrant en σ,

$$\begin{aligned} p(\mu \mid D) &\propto \int_0^\infty \sigma^{-(n+1)} \exp\left[-\frac{n}{2\sigma^2}[s^2 + (\overline{x} - \mu)^2] \right] d\sigma \\ &\propto [s^2 + (\overline{x} - \mu)^2]^{-n/2}\,, \end{aligned} \tag{14}$$

reconeixem que és un nucli de la densitat d'Student $\mathrm{St}(\mu \mid \overline{x}, s/\sqrt{n-1}, n-1)$. Anàlogament, integrant en μ,

$$\begin{aligned} p(\sigma \mid D) &\propto \int_{-\infty}^\infty \sigma^{-(n+1)} \exp\left[-\frac{n}{2\sigma^2}[s^2 + (\overline{x} - \mu)^2] \right] d\mu \\ &\propto \sigma^{-n} \exp\left[-\frac{ns^2}{2\sigma^2} \right]. \end{aligned} \tag{15}$$

Fent un canvi de variables $\lambda = \sigma^{-2}$ s'obté $p(\lambda \mid D) \propto \lambda^{(n-3)/2} e^{ns^2\lambda/2}$, un nucli de la densitat Gamma $\mathrm{Ga}(\lambda \mid (n-1)/2, ns^2/2)$. En termes de la desviació típica σ la densitat final de probabilitat esdevé $p(\sigma \mid D) = p(\lambda \mid D)|\partial\lambda/\partial\sigma| = 2\sigma^{-3}\,\mathrm{Ga}(\sigma^{-2} \mid (n-1)/2, ns^2/2)$, la densitat de la inversa de l'arrel quadrada d'una gamma.

Un exemple freqüent en aquest context el proporcionen mesures de laboratori fetes en condicions de manera que les hipòtesis del teorema central del límit es compleixen, això és, (suposant que no hi ha biaix experimental) les mesures poden ser tractades com una mostra aleatòria d'una distribució normal centrada en la quantitat μ que es mesura, i amb una desviació típica (desconeguda) σ. Suposem, per exemple, que en una classe de física elemental s'experimenta per a mesurar el camp gravitatori g amb un pèndol i un estudiant ha obtingut $n = 20$ mesures de g que proporcionen (en $\mathrm{m/s^2}$) una mitjana

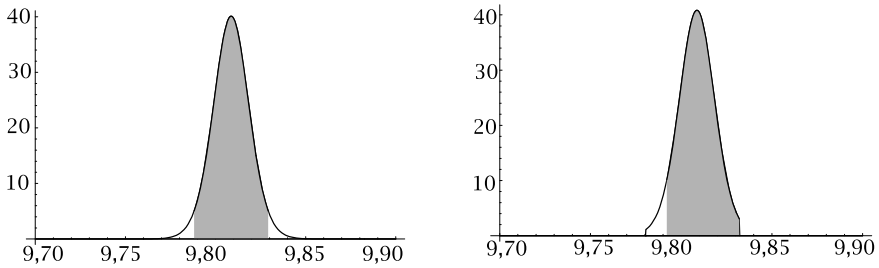

FIGURA 3: Densitat final $p(g \mid m, s, n)$ del valor g del camp gravitatori, donades $n = 20$ mesures normals amb mitjana m =9,8087 i desviació típica $s = 0{,}0428$, *a*) sense informació addicional, i *b*) amb g restringida a $G_c = \{g; 9{,}7803 < g < 9{,}8322\}$. Les àrees ombrejades representen 95 %-regions de confiança de g.

de $\overline{x} = 9{,}8087$, i una desviació típica $s = 0{,}0428$. Sense més informació, la distribució final corresponent és $p(g \mid D) = \mathrm{St}(g \mid 9{,}8087, 0{,}0098, 19)$ representada a la figura 3 *a*). En particular, $\Pr(9{,}788 < g < 9{,}829 \mid D) = 0{,}95$, i per tant, amb la informació proporcionada per aquest experiment, pot esperar-se que el camp gravitatori en el lloc del laboratori estigui entre 9,788 i 9,829 amb probabilitat 0,95.

Formalment, la distribució final de g ha de ser restringida a $g > 0$; però, com és clar a partir de la figura 3 *a*), això no tindrà cap efecte apreciable, atès que la funció de versemblança està concentrada en els valors de g positius.

Suposem ara que s'ensenya a l'estudiant a incorporar a l'anàlisi el fet que sabem que el valor del camp gravitatori g al laboratori està entre 9,7803 m/s^2 (valor mitjà a l'Equador) i 9,8322 m/s^2 (valor mitjà als pols). La distribució final actualitzada serà

$$p(g \mid D, g \in G_c) = \frac{\mathrm{St}(g \mid m, s/\sqrt{n-1}, n)}{\int_{g \in G_c} \mathrm{St}(g \mid m, s/\sqrt{n-1}, n)}, \quad g \in G_c, \tag{16}$$

representada a la figura 3 *b*), on $G_c = \{g; 9{,}7803 < g < 9{,}8322\}$. Es pot integrar numèricament per a comprovar que $\Pr(g > 9{,}792 \mid D, g \in G_c) = 0{,}95$. Si a més també volem inferències sobre la desviació típica σ del procediment de mesura, la distribució final corresponent és $p(\sigma \mid D) = 2\sigma^{-3}\,\mathrm{Ga}(\sigma^{-2} \mid 9{,}5, 0{,}0183)$, que té mitjana $\mathrm{E}[\sigma \mid D] = 0{,}0458$, i dóna $\Pr(0{,}0334 < \sigma < 0{,}0642 \mid D) = 0{,}95$.

3.2 Distribucions predictives

Sigui $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $\boldsymbol{x}_i \in X$ un conjunt d'observacions intercanviables i considerem ara una situació on volem predir el valor d'una observació fu-

tura $\boldsymbol{x} \in X$ generada pel mateix mecanisme aleatori que ha generat les dades D. Dels arguments discutits a la secció 2 es dedueix que la solució a aquest problema de predicció està senzillament encapsulada en la distribució *predictiva* $p(\boldsymbol{x} \mid D)$ que descriu la incertesa del valor que $\boldsymbol{x}$ prendrà, atesa la informació proporcionada per D i qualsevol altre coneixement disponible. Suposem que la informació del context suggereix la suposició que les dades D poden ser considerades com una mostra aleatòria d'una distribució de la família $\{p(\boldsymbol{x} \mid \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, i sigui $p(\boldsymbol{\omega})$ una distribució inicial que descriu la informació disponible sobre el valor de $\boldsymbol{\omega}$. Atès que $p(\boldsymbol{x} \mid \boldsymbol{\omega}, D) = p(\boldsymbol{x} \mid \boldsymbol{\omega})$, es dedueix de càlculs estàndard que $p(\boldsymbol{x} \mid D) = \int_\Omega p(\boldsymbol{x} \mid \boldsymbol{\omega})\, p(\boldsymbol{\omega} \mid D)\, d\boldsymbol{\omega}$, que és una mitjana de les distribucions de probabilitat de $\boldsymbol{x}$ condicionades pel valor (desconegut) de $\boldsymbol{\omega}$, ponderats per la distribució final de $\boldsymbol{\omega}$ suposat D.

Si les hipòtesis sobre el model de probabilitat són correctes, la distribució final predictiva $p(\boldsymbol{x} \mid D)$ convergirà, quan la grandària mostral creixi, a la distribució $p(\boldsymbol{x} \mid \boldsymbol{\omega})$ que ha generat les dades. De fet, la millor tècnica per a analitzar la qualitat de les inferències sobre $\boldsymbol{\omega}$ encapsulades en $p(\boldsymbol{\omega} \mid D)$ és contraposar les dades observades amb la distribució predictiva $p(\boldsymbol{x} \mid D)$ generada per $p(\boldsymbol{\omega} \mid D)$.

4 EXEMPLE *(Predicció en un procés de Poisson.)* Sigui $D = \{r_1, \ldots, r_n\}$ una mostra aleatòria d'una distribució de Poisson $\mathrm{Pn}(r \mid \lambda)$ de paràmetre λ, és a dir, $p(D \mid \lambda) \propto \lambda^t e^{-\lambda n}$, on $t = \sum r_i$. L'absència d'informació inicial (vegeu la secció 5) sobre el valor de λ pot ser descrita formalment amb la funció inicial (impròpia) $p(\lambda) = \lambda^{-1/2}$. D'acord amb el teorema de Bayes, la distribució final corresponent és

$$p(\lambda \mid D) \propto \lambda^t e^{-\lambda n}\, \lambda^{-1/2} \propto \lambda^{t-1/2} e^{-\lambda n}\,, \tag{17}$$

el nucli d'una densitat Gamma $\mathrm{Ga}(\lambda \mid t + 1/2, n)$, amb mitjana $(t + 1/2)/n$. La distribució predictiva corresponent és una mixtura Poisson-Gamma

$$\begin{aligned} p(r \mid D) &= \int_0^\infty \mathrm{Pn}(r \mid \lambda)\ \mathrm{Ga}(\lambda \mid t + \frac{1}{2}, n)\, d\lambda \qquad (18) \\ &= \frac{n^{t+1/2}}{\Gamma(t + 1/2)}\, \frac{1}{r!}\, \frac{\Gamma(r + t + 1/2)}{(1 + n)^{r+t+1/2}}\ . \end{aligned}$$

Suposem, per exemple, que en una empresa que produeix frens per a cotxes, no hi ha hagut cap queixa entre els clients sobre la producció completa en deu mesos consecutius. Sense informació addicional sobre el nombre mitjà λ de queixes per mes, el departament de control de qualitat de l'empresa pot informar que la probabilitat de rebre r queixes sobre la producció del proper mes ve donada per l'equació (18), amb $t = 0$ i $n = 10$. En particular, $p(r = 0 \mid D) = 0{,}953$, $p(r = 1 \mid D) = 0{,}043$, i $p(r = 2 \mid D) = 0{,}003$. Moltes altres situacions poden ser descrites amb el mateix model. Per exemple, si les condicions meteorològiques es mantenen similars en una àrea en concret,

$p(r = 0 \mid D) = 0{,}953$ descriurà la probabilitat que no hi hagi cap inundació el proper any, atesos 10 anys sense inundacions en aquella àrea.

5 EXEMPLE *(Predicció en un procés normal.)* Considerem ara la predicció d'una variable contínua. Sigui $D = \{x_1, \ldots, x_n\}$ una mostra aleatòria d'una distribució normal $N(x \mid \mu, \sigma)$. Tal com dèiem a l'exemple 3, en absència d'informació inicial, la informació sobre els valors de μ i σ es descriu formalment amb la funció inicial *impròpia* $p(\mu, \sigma) = \sigma^{-1}$, que proporciona la densitat final (13). La distribució (final) predictiva corresponent és

$$\begin{aligned} p(x \mid D) &= \int_0^\infty \int_{-\infty}^\infty \mathrm{N}(x \mid \mu, \sigma)\, p(\mu, \sigma \mid D)\, d\mu d\sigma \\ &= \mathrm{St}(x \mid \overline{x}, s\sqrt{\frac{n+1}{n-1}}, n-1)\,. \end{aligned} \tag{19}$$

Si sabem que μ és positiva, la funció inicial adient estarà restringida a

$$p(\mu, \sigma) = \begin{cases} \sigma^{-1}, & \text{si } \mu > 0, \\ 0, & \text{en cas contrari}\,. \end{cases} \tag{20}$$

Però el resultat en l'equació (19) encara serà cert, si resulta que $p(D \mid \mu, \sigma)$, la funció de versemblança, està concentrada en valors positius de μ. Suposem, per exemple, que en l'empresa que produïa cinturons de seguretat per a automòbils, la força de trencament de $n = 10$ cinturons escollits a l'atzar té mitjana $\overline{x} = 28.011\,\mathrm{kN}$ i desviació típica $s = 0{,}443\,\mathrm{kN}$, i que l'especificació d'enginyeria demana tensions de trencament més grans de $26\,\mathrm{kN}$. Si podem assumir que les dades són una mostra aleatòria d'una distribució normal, la funció de versemblança només és rellevant per a valors positius de μ, i si sols es pot utilitzar la informació proporcionada per aquesta mostra, llavors l'enginyer de qualitat pot afirmar que la probabilitat que un cinturó de seguretat escollit a l'atzar del mateix lot que la mostra provada compleixi l'especificació demanada és $\Pr(x > 26 \mid D) = 0{,}9987$. A més, si les condicions de producció es mantenen constants, es pot esperar que el 99,87% dels cinturons tinguin tensions de trencament acceptables.

3.3 Comportament asimptòtic

Estudiarem ara el comportament de les distribucions finals quan la mida de la mostra és gran. Això és important per, almenys, dues raons diferents: (i) els resultats asimptòtics proporcionen aproximacions de primer ordre útils quan les mostres reals són relativament grans, i (ii) les distribucions objectives inicials depenen de les propietats asimptòtiques del model. Sigui $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $\boldsymbol{x} \in X$, una mostra aleatòria de grandària n de $\{p(\boldsymbol{x} \mid \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$. Pot demostrar-se que, quan $n \to \infty$, la distribució final $p(\boldsymbol{\omega} \mid D)$ d'un paràmetre *discret* $\boldsymbol{\omega}$ normalment convergeix a una distribució degenerada que

dóna probabilitat u al valor veritable de $\boldsymbol{\omega}$, i que la distribució final d'un paràmetre *continu* $\boldsymbol{\omega}$ normalment convergeix a una distribució normal centrada a l'*estimació del màxim de versemblança* $\widehat{\boldsymbol{\omega}}$ (abreujat EMV), amb una matriu de variàncies que decreix amb n com $1/n$.

Considerem primer la situació on $\Omega = \{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots\}$ consisteix en un conjunt de valors *numerable* (possiblement infinit), tal que el model de probabilitat que correspon al valor vertader del paràmetre $\boldsymbol{\omega}_t$ és *distingible* dels altres en el sentit que la divergència logarítmica $\delta\{p(\boldsymbol{x} \mid \boldsymbol{\omega}_i) \mid p(\boldsymbol{x} \mid \boldsymbol{\omega}_t)\}$ de cadascun de $p(\boldsymbol{x} \mid \boldsymbol{\omega}_i)$ a $p(\boldsymbol{x} \mid \boldsymbol{\omega}_t)$ és estrictament positiva. Prenent logaritmes en el teorema de Bayes, posant $z_j = \log[p(\boldsymbol{x}_j \mid \boldsymbol{\omega}_i)/p(\boldsymbol{x}_j \mid \boldsymbol{\omega}_t)]$, $j = 1, \ldots, n$, i a partir de la llei forta dels grans nombres per a n quantitats aleatòries condicionalment independents i idènticament distribuïdes $z_1, \ldots, z_n$, es pot demostrar que

$$\lim_{n\to\infty} p(\boldsymbol{\omega}_t \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = 1, \qquad \lim_{n\to\infty} p(\boldsymbol{\omega}_i \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = 0, \quad i \neq t\,. \tag{21}$$

Així, sota condicions de regularitat adients, la probabilitat final del valor vertader del paràmetre convergeix a u quan la mida de la mostra creix.

Considerem ara la situació on $\boldsymbol{\omega}$ és un paràmetre k-dimensional *continu*. Expressant el teorema de Bayes com $p(\boldsymbol{\omega} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \propto \exp\{\log[p(\boldsymbol{\omega})] + \sum_{j=1}^{n} \log[p(\boldsymbol{x}_j \mid \boldsymbol{\omega})]\}$, expandint $\sum_j \log[p(\boldsymbol{x}_j \mid \boldsymbol{\omega})]$ sobre el seu màxim (l'EMV $\widehat{\boldsymbol{\omega}}$), i suposant condicions de regularitat (per assegurar que podem ignorar els termes d'ordre superior al quadràtic i que la suma dels termes obtinguts a partir de la versemblança dominaran el terme que ve de la inicial) es troba que la densitat final de $\boldsymbol{\omega}$ és aproximadament una normal k-variada

$$\begin{aligned} p(\boldsymbol{\omega} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) &\approx \mathrm{N}_k\{\widehat{\boldsymbol{\omega}}, \boldsymbol{S}(D, \widehat{\boldsymbol{\omega}})\}\,, \\ \boldsymbol{S}^{-1}(D, \boldsymbol{\omega}) &= \Big(-\sum_{l=1}^{n} \frac{\partial^2 \log[p(\boldsymbol{x}_l \mid \boldsymbol{\omega})]}{\partial\omega_i \partial\omega_j} \Big)\,. \end{aligned} \tag{22}$$

Una aproximació simple, encara que menys precisa, es pot obtenir usant la llei forta dels grans nombres per a les sumes (22) per a demostrar que $\boldsymbol{S}^{-1}(D, \widehat{\boldsymbol{\omega}}) \approx n\,\boldsymbol{F}(\widehat{\boldsymbol{\omega}})$, on $\boldsymbol{F}(\boldsymbol{\omega})$ és *la matriu d'informació de Fisher*, d'element genèric

$$\boldsymbol{F}_{ij}(\boldsymbol{\omega}) = -\int_X p(\boldsymbol{x} \mid \boldsymbol{\omega}) \; \frac{\partial^2 \log[p(\boldsymbol{x} \mid \boldsymbol{\omega})]}{\partial\omega_i \partial\omega_j}\, d\boldsymbol{x}\,, \tag{23}$$

és a dir,

$$p(\boldsymbol{\omega} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \approx N_k(\boldsymbol{\omega} \mid \widehat{\boldsymbol{\omega}}, n^{-1}\,\boldsymbol{F}^{-1}(\widehat{\boldsymbol{\omega}}))\,. \tag{24}$$

Així, sota condicions de regularitat adients, la densitat final de probabilitat del paràmetre vectorial $\boldsymbol{\omega}$ s'aproxima, quan la grandària mostral creix, a una densitat normal multivariant centrada en l'EMV $\widehat{\boldsymbol{\omega}}$, amb una matriu de variàncies que decreix amb n i amb n^{-1}.

2 EXEMPLE *(Inferència sobre un paràmetre binomial, continuació.)* Siguin $D = (x_1, \ldots, x_n)$, n proves independents de Bernoulli amb paràmetre θ, de manera que $p(D \mid \theta, n) = \theta^r (1-\theta)^{n-r}$. El màxim d'aquesta funció de versemblança es troba a $\hat{\theta} = r/n$, i la funció d'informació de Fisher és $F(\theta) = \theta^{-1}(1-\theta)^{-1}$. Així, dels resultats anteriors, la distribució final de θ serà aproximadament normal,

$$p(\theta \mid r, n) \approx \mathrm{N}(\theta \mid \hat{\theta}, s(\hat{\theta})/\sqrt{n}), \qquad s(\theta) = \{\theta(1-\theta)\}^{1/2} \tag{25}$$

amb mitjana $\hat{\theta} = r/n$ i variància $\hat{\theta}(1-\hat{\theta})/n$. Això proporciona una aproximació raonable a la final exacta si *a*) la inicial $p(\theta)$ és relativament «plana» en la regió on la funció de versemblança importa, i *b*) ambdós r i n són moderadament grans. Si, posem, $n = 1.500$ i $r = 720$, això porta a $p(\theta \mid D) \approx \mathrm{N}(\theta \mid 0{,}480, 0{,}013)$ i $\Pr(\theta > 0{,}5 \mid D) \approx 0{,}940$, que pot ser comparat amb el valor exacte $\Pr(\theta > 0{,}5 \mid D) = 0{,}933$ obtingut a partir de la distribució final que correspon a la inicial $\mathrm{Be}(\theta \mid 50, 50)$.

Del comportament asimptòtic de la final *conjunta* d'$\boldsymbol{\omega}$ i de les propietats de la distribució normal multivariant es dedueix que si es descompon el paràmetre vectorial en $\boldsymbol{\omega} = (\boldsymbol{\theta}, \boldsymbol{\lambda})$ i es parteix la matriu d'informació de Fisher de la manera corresponent, és a dir,

$$\boldsymbol{F}(\boldsymbol{\omega}) = \boldsymbol{F}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \begin{pmatrix} \boldsymbol{F}_{\theta\theta}(\boldsymbol{\theta}, \boldsymbol{\lambda}) & \boldsymbol{F}_{\theta\lambda}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ \boldsymbol{F}_{\lambda\theta}(\boldsymbol{\theta}, \boldsymbol{\lambda}) & \boldsymbol{F}_{\lambda\lambda}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \end{pmatrix} \tag{26}$$

i

$$\boldsymbol{S}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \boldsymbol{F}^{-1}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \begin{pmatrix} \boldsymbol{S}_{\theta\theta}(\boldsymbol{\theta}, \boldsymbol{\lambda}) & \boldsymbol{S}_{\theta\lambda}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ \boldsymbol{S}_{\lambda\theta}(\boldsymbol{\theta}, \boldsymbol{\lambda}) & \boldsymbol{S}_{\lambda\lambda}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \end{pmatrix}, \tag{27}$$

llavors la distribució final *marginal* de $\boldsymbol{\theta}$ és

$$p(\boldsymbol{\theta} \mid D) \approx \mathrm{N}\{\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}, n^{-1} \boldsymbol{S}_{\theta\theta}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\lambda}})\}, \tag{28}$$

i la distribució final *condicional* de $\boldsymbol{\lambda}$ donat $\boldsymbol{\theta}$ és

$$p(\boldsymbol{\lambda} \mid \boldsymbol{\theta}, D) \approx \mathrm{N}\{\boldsymbol{\lambda} \mid \hat{\boldsymbol{\lambda}} - \boldsymbol{F}_{\lambda\lambda}^{-1}(\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}) \boldsymbol{F}_{\lambda\theta}(\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}),\ n^{-1} \boldsymbol{F}_{\lambda\lambda}^{-1}(\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}})\}. \tag{29}$$

Noteu que $\boldsymbol{F}_{\lambda\lambda}^{-1} = \boldsymbol{S}_{\lambda\lambda}$ si (i només si) $\boldsymbol{F}$ és diagonal en blocs, *i.e.*, si (i només si) $\boldsymbol{\theta}$ i $\boldsymbol{\lambda}$ són asimptòticament independents.

3 EXEMPLE *(Inferència sobre paràmetres normals, continuació.)* Sigui $D = (x_1, \ldots, x_n)$ una mostra aleatòria d'una distribució normal $\mathrm{N}(x \mid \mu, \sigma)$. La funció de versemblança corresponent $p(D \mid \mu, \sigma)$ assoleix el màxim a $(\hat{\mu}, \hat{\sigma}) = (\overline{x}, s)$, i la matriu d'informació de Fisher és diagonal, amb $F_{\mu\mu} = \sigma^{-2}$. D'aquí, la distribució final de μ és *aproximadament* $\mathrm{N}(\mu \mid \overline{x}, s/\sqrt{n})$; això pot ser comparat amb el resultat *exacte* $p(\mu \mid D) = \mathrm{St}(\mu \mid \overline{x}, s/\sqrt{n-1}, n-1)$ obtingut prèviament sota la suposició de no-coneixement inicial.

4 Sumaris d'inferència

Des del punt de vista bayesià, el resultat final d'un problema d'inferència sobre *qualsevol* quantitat desconeguda és precisament la distribució final corresponent. Així, donades unes dades D i unes condicions C, *tot* el que pot ser dit sobre una funció arbitrària $\boldsymbol{\omega}$ dels paràmetres del model està contingut en la distribució final $p(\boldsymbol{\omega} \mid D, C)$, i *tot* el que pot ser dit sobre qualsevol funció $\boldsymbol{y}$ d'observacions futures a partir del mateix model està contingut en la seva distribució final predictiva $p(\boldsymbol{y} \mid D, C)$. Tal com dèiem abans, la inferència bayesiana es pot descriure tècnicament com un problema de decisió on l'espai d'accions disponibles és la classe de les distribucions de probabilitat final de la quantitat d'interès que són compatibles amb les hipòtesis acceptades.

Però per a facilitar a l'usuari l'assimilació de les conclusions, és sovint convenient *resumir* la informació continguda en la distribució final per *a*) donar valors de la quantitat d'interès que, a la vista de les dades, és raonable que siguin «properes» al seu valor veritable i per a *b*) mesurar la compatibilitat dels resultats amb valors hipotètics de la quantitat d'interès que poden haver estat suggerits en el context de la investigació. En aquesta secció, considerarem breument les contraparts bayesianes dels problemes tradicionals d'*estimació* i *test d'hipòtesi*.

4.1 Estimació

En una o dues dimensions, un gràfic de la densitat de probabilitat final de la quantitat d'interès (o de la funció de repartiment de massa en el cas discret) proporciona immediatament un resum «impressionista» intuïtiu de les conclusions més importants que possiblement podran ser extretes de les dades. De fet, aquest és un punt molt apreciat pels usuaris, i pot ser citat com un avantatge important dels mètodes bayesians. A partir d'un gràfic de la densitat final, es distingeix de manera fàcil la regió on és probable que es trobi una quantitat d'interès univariant (donades les dades). Per exemple, totes les conclusions importants sobre el valor del camp gravitatori de l'exemple 3 són qualitativament observables a la figura 3. Però això no es pot estendre fàcilment en més de dues dimensions i, a més, normalment es desitgen conclusions *quantitatives* (en una forma més simple que la proporcionada per l'expressió matemàtica de la distribució final).

Estimació puntual. Siguin D les dades disponibles, que suposarem que han estat generades per a un model de probabilitat $\{p(D \mid \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, i sigui $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \Theta$ la quantitat d'interès. Un *estimador puntual* de $\boldsymbol{\theta}$ és una funció de les dades $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(D)$ que pot ser considerada un substitut adient al valor real, desconegut, de $\boldsymbol{\theta}$. Formalment, per a escollir un estimador puntual per a $\boldsymbol{\theta}$ és planteja un *problema de decisió*, on l'espai d'accions és la classe Θ dels valors possibles $\boldsymbol{\theta}$. Des d'una perspectiva de teoria de la decisió, escollir un estimador puntual $\tilde{\boldsymbol{\theta}}$ d'una quantitat $\boldsymbol{\theta}$ no és afirmar alguna cosa sobre el valor

de $\boldsymbol{\theta}$ sinó que és la *decisió* d'actuar com si penséssim que $\widetilde{\boldsymbol{\theta}}$ és $\boldsymbol{\theta}$ (malgrat que el desig d'afirmar alguna cosa simple pot ser també la raó per a obtenir un estimador). Tal com està prescrit pels fonaments de la teoria de la decisió (secció 2), per a resoldre aquest problema de decisió és necessari especificar una *funció de pèrdua* $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$ que mesuri les conseqüències d'actuar *com si* el valor veritable de la quantitat d'interès fos $\widetilde{\boldsymbol{\theta}}$, quan de fet és $\boldsymbol{\theta}$. Quan s'utilitza $\widetilde{\boldsymbol{\theta}}$ la pèrdua esperada final és

$$\overline{L}[\widetilde{\boldsymbol{\theta}} \mid D] = \int_{\Theta} L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}\,, \tag{30}$$

i el corresponent *estimador de Bayes* $\boldsymbol{\theta}^*$ és la funció de les dades, $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(D)$, que minimitza aquesta esperança.

6 Exemple *(Estimadors de Bayes convencionals.)* Per a un model i unes dades donades, l'estimador de Bayes òbviament depèn de la funció de pèrdua escollida. La funció de pèrdua és específica del context i ha de ser escollida en termes de l'ús previst de l'estimació; però quan no es preveuen utilitzacions particulars, s'han proposat diferents funcions de pèrdua convencionals. Aquestes funcions produeixen estimadors que poden mirar com simples descripcions de la *posició* de la distribució final. Per exemple, si la funció de pèrdua és quadràtica, és a dir, $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^t(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$; llavors l'estimador de Bayes és la *mitjana final* $\boldsymbol{\theta}^* = E[\boldsymbol{\theta} \mid D]$, suposant que la mitjana existeix. Anàlogament, si la funció de pèrdua és una funció zero-u, és a dir, $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 0$ si $\widetilde{\boldsymbol{\theta}}$ pertany a una bola de radi ϵ centrada en $\boldsymbol{\theta}$ i $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 1$ en cas contrari, llavors l'estimador de Bayes $\boldsymbol{\theta}^*$ tendeix a la *moda final* quan el radi ϵ tendeix a zero, suposant que existeix una única moda. Si θ és univariant i la funció de pèrdua és lineal, és a dir, $L(\widetilde{\theta}, \theta) = c_1(\widetilde{\theta} - \theta)$ si $\widetilde{\theta} \geq \theta$, i $L(\widetilde{\theta}, \theta) = c_2(\theta - \widetilde{\theta})$ altrament, llavors l'estimador de Bayes és el *quantil final* d'ordre $c_2/(c_1 + c_2)$, és a dir, $\Pr[\theta < \theta^*] = c_2/(c_1 + c_2)$. En particular, si $c_1 = c_2$, l'estimador de Bayes és la *mediana final*. Els resultats obtinguts per a les funcions de pèrdua lineals il·lustren clàrament el fet que *qualsevol* possible valor del paràmetre pot esdevenir un estimador de Bayes: tot depèn de la funció de pèrdua que descrigui les conseqüències dels usos previstos de les estimacions.

7 Exemple *(Estimació intrínseca.)* Les funcions de pèrdua convencionals típicament no són invariants sota reparametritzacions, això és, l'estimador de Bayes $\boldsymbol{\phi}^*$ d'una transformació bijectiva $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ del paràmetre original $\boldsymbol{\theta}$ no és necessàriament $\boldsymbol{\phi}(\boldsymbol{\theta}^*)$ (la mediana final, que *és* invariant, és una excepció interessant). A més, les funcions de pèrdua convencionals estan enfocades sobre la «distància» entre l'estimació $\widetilde{\boldsymbol{\theta}}$ i el valor veritable $\boldsymbol{\theta}$, i no en la «distància» entre els models probabilístics que retolen. Les pèrdues intrínseques es concentren directament en com de diferent és el model de probabilitat $p(D \mid \boldsymbol{\theta}, \boldsymbol{\lambda})$ de la seva aproximació més propera dintre la família $\{p(D \mid \widetilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}_i), \boldsymbol{\lambda}_i \in \Lambda\}$, i

típicament produeixen solucions invariants. Un exemple interessant és la *discrepància intrínseca*, $d(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$, definida com la mínima divergència logarítmica entre el model de probabilitat retolat per $\boldsymbol{\theta}$ i el model de probabilitat retolat per $\widetilde{\boldsymbol{\theta}}$; quan no hi ha paràmetres marginals, ve donada per

$$d(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \min\{\delta(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta}), \delta(\boldsymbol{\theta} \mid \widetilde{\boldsymbol{\theta}})\}\,, \tag{31}$$

on

$$\delta(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}) = \int_T p(\boldsymbol{t} \mid \boldsymbol{\theta}) \log \frac{p(\boldsymbol{t} \mid \boldsymbol{\theta})}{p(\boldsymbol{t} \mid \boldsymbol{\theta}_i)}\, d\boldsymbol{t}\,,$$

on $\boldsymbol{t} = \boldsymbol{t}(D) \in T$ és *qualsevol* estadístic suficient (que pot ben bé ser tot el conjunt de dades D). La definició s'estén fàcilment a problemes amb paràmetres marginals; en aquest cas,

$$\delta(\boldsymbol{\theta}_i \mid \boldsymbol{\omega}) = \delta(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}, \boldsymbol{\lambda}) = \inf_{\boldsymbol{\lambda}_i \in \Lambda} \int_T p(\boldsymbol{t} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}) \log \frac{p(\boldsymbol{t} \mid \boldsymbol{\theta}, \boldsymbol{\lambda})}{p(\boldsymbol{t} \mid \boldsymbol{\theta}_i, \boldsymbol{\lambda}_i)}\, d\boldsymbol{t} \tag{32}$$

mesura la divergència logarítmica de $p(\boldsymbol{t} \mid \boldsymbol{\theta}, \boldsymbol{\lambda})$ a la seva aproximació més propera amb $\boldsymbol{\theta} = \boldsymbol{\theta}_i$, i la funció de pèrdua $d(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\omega}) = \min\{\delta(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}), \delta(\boldsymbol{\theta} \mid \widetilde{\boldsymbol{\theta}}, \boldsymbol{\lambda})\}$ ara depèn de tot el vector de paràmetres $\boldsymbol{\omega} = (\boldsymbol{\theta}, \boldsymbol{\lambda})$. Malgrat que no es mostra explícitament en la notació, la funció de discrepància intrínseca normalment depèn de la gradària mostral n; de fet, quan les dades són una mostra aleatòria $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ d'un model $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\lambda})$, llavors

$$\delta(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}, \boldsymbol{\lambda}) = n \inf_{\boldsymbol{\lambda}_i \in \Lambda} \int_X p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}) \log \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\lambda})}{p(\boldsymbol{x} \mid \boldsymbol{\theta}_i, \boldsymbol{\lambda}_i)}\, d\boldsymbol{x}\,, \tag{33}$$

de manera que la discrepància associada amb el model complet és simplement n vegades la discrepància corresponent a una observació. La discrepància intrínseca és una funció de pèrdua simètrica i no negativa, amb una interpretació directa en termes de teoria de la informació com la mínima quantitat d'informació que s'espera que sigui necessari per a distingir entre el model $p(D \mid \boldsymbol{\theta}, \boldsymbol{\lambda})$ i la seva aproximació més propera dintre de la classe $\{p(D \mid \widetilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}_i), \boldsymbol{\lambda}_i \in \Lambda\}$. A més, és invariant sota reparametritzacions bijectives del paràmetre d'interès $\boldsymbol{\theta}$, i no depèn de l'elecció del paràmetre marginal $\boldsymbol{\lambda}$. L'*estimador intrínsec* s'obté naturalment minimitzant la discrepància intrínseca esperada final

$$\overline{d}(\widetilde{\boldsymbol{\theta}} \mid D) = \int_\Omega d(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\omega})\, p(\boldsymbol{\omega} \mid D)\, d\boldsymbol{\omega}\,. \tag{34}$$

Atès que la discrepància intrínseca és invariant sota reparametrització, minimitzar la seva esperança final produeix estimadors invariants.

2 EXEMPLE *(Inferència sobre un paràmetre binomial, continuació.)* A l'estimació d'una proporció binomial θ, suposades unes dades $D = (n, r)$ i una inicial Beta $\mathrm{Be}(\theta \mid \alpha, \beta)$, l'estimador de Bayes amb funció de pèrdua quadràtica (la

mitjana final corresponent) és $E[\theta \mid D] = (r+\alpha)/(n+\alpha+\beta)$, mentre que l'estimador amb pèrdua quadràtica de, posem, el *log-odds* $\phi(\theta) = \log[\theta/(1-\theta)]$, és $E[\phi \mid D] = \psi(r+\alpha) - \psi(n-r+\beta)$ (on $\psi(x) = d\log[\Gamma(x)]/dx$ és la funció *digamma*), que *no* és igual a $\phi(E[\theta \mid D])$. La funció de pèrdua intrínseca en aquest problema és

$$d(\widetilde{\theta}, \theta) = n \, \min\{\delta(\widetilde{\theta} \mid \theta), \delta(\theta \mid \widetilde{\theta})\},$$

$$\delta(\theta_i \mid \theta) = \theta \log \frac{\theta}{\theta_i} + (1-\theta) \log \frac{1-\theta}{1-\theta_i} \;, \tag{35}$$

i l'estimador intrínsec corresponent θ^* s'obté minimitzant la pèrdua esperada final $\overline{d}(\widetilde{\theta} \mid D) = \int d(\widetilde{\theta}, \theta)\, p(\theta \mid D)\, d\theta$. El valor exacte de θ^* es pot obtenir per minimització numèrica, però una aproximació molt bona ve donada per

$$\theta^* \approx \frac{1}{2}\,\frac{r+\alpha}{n+\alpha+\beta} + \frac{1}{2}\,\frac{e^{\psi(r+\alpha)}}{e^{\psi(r+\alpha)} + e^{\psi(n-r+\beta)}}\,. \tag{36}$$

Com que l'estimació intrínseca és un procediment invariant, l'estimador intrínsec del *log-odds* serà senzillament el *log-odds* de l'estimador intrínsec de θ. Tal com es podia esperar, quan $r+\alpha$ i $n-r+\beta$ són grans, tots els estimadors de Bayes de qualsevol funció suficientment regular $\phi(\theta)$ seran propers a $\phi(E[\theta \mid D])$.

Estimació per interval. Per a descriure el contingut inferencial de la distribució final de la quantitat d'interès $p(\boldsymbol{\theta} \mid D)$ és sovint convenient proporcionar regions $R \subset \Theta$ de probabilitat donada sota $p(\boldsymbol{\theta} \mid D)$. Per exemple, la identificació de regions que contenen el 50%, 90%, 95% o 99% de probabilitat sota la distribució final pot ser suficient per a transmetre els missatges quantitatius implícits en $p(\boldsymbol{\theta} \mid D)$; de fet, aquesta és la base intuïtiva de les representacions gràfiques de les distribucions univariants com les que proporcionen els *boxplots*. De qualsevol regió $R \subset \Theta$ tal que $\int_R p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta} = q$, és a dir, que donades unes dades D, el valor veritable de $\boldsymbol{\theta}$ pertany a R amb probabilitat q, és diu que és una *regió q-creïble* final de $\boldsymbol{\theta}$. Noteu que això proporciona una afirmació intuïtiva directa i immediata sobre la quantitat d'interès desconegut $\boldsymbol{\theta}$ en termes de probabilitat, en marcat contrast amb les afirmacions circumlocutòries que donen els intervals de confiança freqüentistes. Clarament, per a qualsevol q en general hi ha infinites regions creïbles. Una regió creïble és invariant sota reparametrització; així, per a qualsevol regió q-creïble R de $\boldsymbol{\theta}$, $\boldsymbol{\phi}(R)$ és una regió q-creïble de $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$. Algunes vegades, les regions creïbles se seleccionen de manera que tinguin la mida mínima (llargària, àrea, volum), i s'obtenen regions amb la màxima densitat de probabilitat (MDP), on tots els punts de la regió tenen densitat de probabilitat més gran que tots els punts de fora. Però les regions MDP *no* són invariants sota reparametrització: la imatge $\boldsymbol{\phi}(R)$ d'una regió MDP R serà una regió creïble per a $\boldsymbol{\phi}$, però en general no serà MDP; de fet, no hi ha cap argument convincent per a restringir-nos

a les regions creïbles MDP. Sovint s'utilitzen els quantils finals per a derivar regions creïbles. Així, si $\boldsymbol{\theta}_q = \boldsymbol{\theta}_q(D)$ és el $100q\,\%$ quantil final de $\boldsymbol{\theta}$, llavors $R = \{\boldsymbol{\theta};\, \boldsymbol{\theta} \leq \boldsymbol{\theta}_q\}$ és una regió q-creïble unilateral, normalment única i és invariant sota reparametrització. De fet, és fàcil calcular regions q-creïbles de la forma $R = \{\boldsymbol{\theta};\, \boldsymbol{\theta}_{q/2} \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}_{1-q/2}\}$, i sovint són citades en preferència a regions MDP.

3 EXEMPLE *(Inferència sobre paràmetres normals, continuació.)* En l'exemple numèric sobre el valor del camp gravitatori descrit a la figura 3 *a*), l'interval $[9{,}788, 9{,}829]$ de la densitat final no restringida de g és una regió MDP, 95 %-creïble per g. Anàlogament, l'interval $[9{,}7803, 9{,}8322]$ de la figura 3 *b*) també és una regió 95 %-creïble per a g, però no és MDP.

El concepte de regions creïbles per a funcions $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega})$ del vector de paràmetres s'estén de manera òbvia als problemes de predicció. Així, una regió final q-creïble per a $\boldsymbol{x} \in X$ és un subconjunt R de l'espai de resultats X amb probabilitat predictiva final q, és a dir, amb $\int_R p(\boldsymbol{x} \mid D) d\boldsymbol{x} = q$.

4.2 Test d'hipòtesis

La distribució final $p(\boldsymbol{\theta} \mid D)$ de la quantitat d'interès $\boldsymbol{\theta}$ transmet immediatament una informació intuïtiva sobre els valors de $\boldsymbol{\theta}$ que, donat el model assumit, poden ser acceptats com a *compatibles* amb les dades observades D, és a dir, aquells que tenen densitat de probabilitat relativament alta. Algunes vegades, al llarg de la investigació es fa evident que una *restricció* $\boldsymbol{\theta} \in \Theta_0 \subset \Theta$ dels valors possibles de la quantitat d'interès (on Θ_0 pot contenir, potser, un únic valor $\boldsymbol{\theta_0}$) mereix especial atenció, ja sigui perquè en restringir $\boldsymbol{\theta}$ a Θ_0 simplificarà molt el model, o perquè arguments específics addicionals del context suggereixen que $\boldsymbol{\theta} \in \Theta_0$. Intuïtivament, la *hipòtesi* $H_0 \equiv \{\boldsymbol{\theta} \in \Theta_0\}$ ha de ser jutjada *compatible* amb les dades observades D si hi ha elements en Θ_0 amb una densitat final relativament alta. Però, sovint, es necessita una conclusió més exacta i, un altre cop, això és possible adoptant una perspectiva de teoria de la decisió. Formalment, contrastar la hipòtesi $H_0 \equiv \{\boldsymbol{\theta} \in \Theta_0\}$ és un *problema de decisió* on l'espai d'accions només té dos elements: acceptar (a_0) o rebutjar (a_1) la restricció proposada. Per a resoldre aquest problema de decisió, cal especificar una funció de pèrdua adient, $L(a_i, \boldsymbol{\theta})$, que mesuri les conseqüències d'acceptar o rebutjar H_0 com una funció del valor real $\boldsymbol{\theta}$ del vector d'interès. Noteu que això demana establir una *alternativa* a_1 a acceptar H_0; però això calia esperar-ho, ja que una acció no es pren perquè sigui bona, sinó perquè és millor que qualsevol altra en què es pugui pensar.

Donades unes dades D, l'acció òptima serà rebutjar H_0 si (i només si) la pèrdua final esperada d'acceptar, $\int_\Theta L(a_0, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$, és més gran que la corresponent pèrdua final esperada de rebutjar, $\int_\Theta L(a_1, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$, és a dir, si (i només si)

$$\int_\Theta [L(a_0, \boldsymbol{\theta}) - L(a_1, \boldsymbol{\theta})]\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta} = \int_\Theta \Delta L(\boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta} > 0\,. \qquad (37)$$

Llavors, només cal especificar la diferència de pèrdues $\Delta L(\boldsymbol{\theta}) = L(a_0, \boldsymbol{\theta}) - L(a_1, \boldsymbol{\theta})$, que mesura el *benefici* de rebutjar H_0 com una funció de $\boldsymbol{\theta}$. Així, tal com diu el sentit comú, la hipòtesi H_0 ha de ser rebutjada quan el benefici esperat de rebutjar H_0 és positiu.

Un element crucial en l'especificació de la funció de pèrdua és la descripció de què s'entén per a rebutjar H_0. Per definició, a_0 vol dir actuar *com si* H_0 fos veritat, *i.e.*, com si $\boldsymbol{\theta} \in \Theta_0$, però hi ha almenys dues opcions òbvies per a l'acció alternativa a_1. Aquesta pot voler dir *a*) la *negació* d'H_0, que és actuar com si $\boldsymbol{\theta} \notin \Theta_0$ o, alternativament, pot voler dir *b*) rebutjar la simplificació implicada per H_0 i conservar el model no restringit, $\boldsymbol{\theta} \in \Theta$, que és veritat per hipòtesi. A la bibliografia s'han analitzat les dues opcions, malgrat que es pot argumentar que a l'anàlisi de dades científiques els tests d'hipòtesis més habituals són del segon tipus. De fet, en establir un model, identificat per $H_0 \equiv \{\boldsymbol{\theta} \in \Theta_0\}$, molt sovint se submergeix en un model més general, $\{\boldsymbol{\theta} \in \Theta, \Theta_0 \subset \Theta\}$, construït per incloure desviacions previsibles d'H_0, i es demana comprovar quan les dades actuals disponibles D són encara compatibles amb $\boldsymbol{\theta} \in \Theta_0$, o quan l'extensió a $\boldsymbol{\theta} \in \Theta$ és realment necessària.

8 EXEMPLE *(Test d'hipòtesis convencional.)* Sigui $p(\boldsymbol{\theta} \mid D)$, $\boldsymbol{\theta} \in \Theta$, la distribució final de la quantitat d'interès, sigui a_0 la decisió de treballar sota la restricció $\boldsymbol{\theta} \in \Theta_0$ i sigui a_1 la decisió de treballar sota la restricció complementària $\boldsymbol{\theta} \notin \Theta_0$. Suposem també que l'estructura de pèrdua té la forma simple zero-u donada per $\{L(a_0, \boldsymbol{\theta}) = 0, L(a_1, \boldsymbol{\theta}) = 1\}$ si $\boldsymbol{\theta} \in \Theta_0$ i, anàlogament, $\{L(a_0, \boldsymbol{\theta}) = 1, L(a_1, \boldsymbol{\theta}) = 0\}$ si $\boldsymbol{\theta} \notin \Theta_0$, això és, el *benefici* $\Delta L(\boldsymbol{\theta})$ de rebutjar H_0 és 1 si $\boldsymbol{\theta} \notin \Theta_0$ i és -1 en cas contrari. Amb aquesta funció de pèrdua és immediat que l'acció òptima és rebutjar H_0 si (i només si) $\Pr(\boldsymbol{\theta} \notin \Theta_0 \mid D) > \Pr(\boldsymbol{\theta} \in \Theta_0 \mid D)$. Noteu que aquesta formulació demana que $\Pr(\boldsymbol{\theta} \in \Theta_0) > 0$, és a dir, que la hipòtesi H_0 tingui probabilitat inicial estrictament positiva. Si $\boldsymbol{\theta}$ és un paràmetre continu i Θ_0 té mesura zero (per exemple si H_0 consisteix en un punt $\boldsymbol{\theta_0}$), llavors cal utilitzar distribucions inicials no regulars «puntuals» que concentrin una massa de probabilitat positiva en $\boldsymbol{\theta}_0$.

9 EXEMPLE *(Contrast d'hipòtesis intrínsec.)* Un altre cop, sigui $p(\boldsymbol{\theta} \mid D)$, $\boldsymbol{\theta} \in \Theta$, la distribució final de la quantitat d'interès, i sigui a_0 la decisió de treballar sota la restricció $\boldsymbol{\theta} \in \Theta_0$, però sigui ara a_1 la decisió de conservar el model general no restringit $\boldsymbol{\theta} \in \Theta$. En aquest cas, podem suposar sense risc que el benefici $\Delta L(\boldsymbol{\theta})$ de rebutjar H_0 com a funció de $\boldsymbol{\theta}$ té la forma $\Delta L(\boldsymbol{\theta}) = d(\Theta_0, \boldsymbol{\theta}) - d^*$, per a algun $d^* > 0$, on *a*) $d(\Theta_0, \boldsymbol{\theta})$ és una mesura de la discrepància entre el model assumit $p(D \mid \boldsymbol{\theta})$ i la seva aproximació més propera dintre de la classe $\{p(D \mid \boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \in \Theta_0\}$, de manera que $d(\Theta_0, \boldsymbol{\theta}) = 0$ quan $\boldsymbol{\theta} \in \Theta_0$, i *b*) d^* és una *constant d'utilitat* que depèn del context, que mesura el benefici (necessàriament positiu) de poder treballar amb el model més simple quan és cert. A continuació descriurem l'elecció d'ambdues $d(\Theta_0, \boldsymbol{\theta})$ i d^* que poden ser adients per a ús general.

Per raons similars a les esgrimides en estimació puntual, una elecció atractiva per a la funció $d(\Theta_0, \boldsymbol{\theta})$ és una extensió adient de la discrepància intrínseca; quan no hi ha paràmetres marginals, s'expressa amb

$$d(\Theta_0, \boldsymbol{\theta}) = \min\{\inf_{\boldsymbol{\theta}_0 \in \Theta_0} \delta(\boldsymbol{\theta}_0 \mid \boldsymbol{\theta}), \inf_{\boldsymbol{\theta}_0 \in \Theta_0} \delta(\boldsymbol{\theta} \mid \boldsymbol{\theta}_0)\}, \tag{38}$$

on $\delta(\boldsymbol{\theta}_0 \mid \boldsymbol{\theta}) = \int_T p(\boldsymbol{t} \mid \boldsymbol{\theta}) \log\{p(\boldsymbol{t} \mid \boldsymbol{\theta})/p(\boldsymbol{t} \mid \boldsymbol{\theta}_0)\} d\boldsymbol{t}$, i $\boldsymbol{t} = \boldsymbol{t}(D) \in T$ és *qualsevol* estadístic suficient, que pot ser tot el conjunt de dades D. Com abans, si les dades $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ consisteixen en una mostra de $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, llavors

$$\delta(\boldsymbol{\theta}_0 \mid \boldsymbol{\theta}) = n \int_X p(\boldsymbol{x} \mid \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta})}{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0)} \, d\boldsymbol{x}. \tag{39}$$

Naturalment, la funció $d(\Theta_0, \boldsymbol{\theta})$ es redueix a la discrepància intrínseca $d(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ de l'exemple 6 quan Θ_0 conté només un element $\boldsymbol{\theta}_0$. A més, com en el cas de l'estimació, la definició s'estén fàcilment als problemes amb paràmetres marginals, amb

$$\delta(\boldsymbol{\theta}_0 \mid \boldsymbol{\theta}, \boldsymbol{\lambda}) = \inf_{\lambda_0 \in \Lambda} \int_T p(\boldsymbol{t} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}) \log \frac{p(\boldsymbol{t} \mid \boldsymbol{\theta}, \boldsymbol{\lambda})}{p(\boldsymbol{t} \mid \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0)} \, d\boldsymbol{t}. \tag{40}$$

La hipòtesi H_0 ha de ser rebutjada si el benefici final esperat de rebutjar és

$$\overline{d}(\Theta_0, D) = \int_\Theta d(\Theta_0, \boldsymbol{\theta}) \, p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta} > d^*, \tag{41}$$

per a algun $d^* > 0$. Es comprova fàcilment que la funció $\overline{d}(\Theta_0, D)$ és no negativa. A més, si $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ és una transformació bijectiva de $\boldsymbol{\theta}$, llavors $\overline{d}(\boldsymbol{\phi}(\Theta_0), D) = \overline{d}(\Theta_0, D)$, és a dir, la pèrdua intrínseca de rebutjar H_0 és invariant sota reparametritzacions.

Es pot demostrar que, quan la grandària mostral creix, el valor esperat de $\overline{d}(\Theta_0, D)$ tendeix a 1 quan H_0 és veritat, i tendeix a infinit en cas contrari; així podem mirar $\overline{d}(\Theta_0, D)$ com una mesura positiva contínua, de com d'inadequat (en unitats de pèrdua d'informació) és simplificar el model acceptant H_0. En llenguatge tradicional, $\overline{d}(\Theta_0, D)$ és un *test estadístic* per a H_0 i la hipòtesi ha de ser rebutjada si el valor de $\overline{d}(\Theta_0, D)$ sobrepassa algun *valor crític* d^*. En forta disparitat amb els tests d'hipòtesis convencionals, aquest valor crític d^* és una constant d'utilitat positiva d^*, dependent del context, que pot ser descrita amb precisió com el nombre d'*unitats d'informació* que qui pren la decisió pot perdre per a poder treballar amb el model més simple H_0, i no depèn de les propietats mostrals del model de probabilitat. El procediment pot ser usat amb distribucions inicials regulars contínues estàndards, fins i tot en tests d'hipòtesis *puntuals*, on Θ_0 és un conjunt de mesura zero (com és el cas quan $\boldsymbol{\theta}$ és continu i Θ_0 conté només un punt $\boldsymbol{\theta}_0$). Naturalment, per a implementar el test cal escollir la constant d'utilitat d^* que defineix la regió de rebuig.

Totes les mesures es basen en la comparació amb un estàndard; la comparació amb el problema «canònic» de contrastar un valor $\mu = \mu_0$ per a la mitjana

de la distribució normal amb variància coneguda (vegeu més endavant) fa possible *calibrar l'escala d'informació*. Els valors de $\overline{d}(\Theta_0, D)$ per sobre d'1 han de ser mirats com una indicació de no-evidència en contra d'H_0, ja que el valor esperat de $\overline{d}(\Theta_0, D)$ sota H_0 és precisament u. Els valors de $\overline{d}(\Theta_0, D)$ per sobre de 2,5 i 5 (respectivament) han de ser considerats com una indicació de lleugera evidència contra H_0 i evidència significativa contra H_0 (respectivament) ja que, en el problema normal canònic, aquests valors corresponen a la mitjana mostral observada $\overline{x}$ respectivament a distància 2 o 3 desviacions estàndards finals del valor nul μ_0. Noteu que, en aguda diferència amb els tests d'hipòtesis freqüentistes, on es recomana confusament ajustar el nivell de significació per dimensionalitat i grandària mostral, aquí es proporciona una escala absoluta (en unitats d'informació) que continua essent vàlida per a qualsevol grandària mostral i qualsevol dimensionalitat.

10 Exemple *(Test sobre el valor de la mitjana normal.)* Suposem que les dades $D = \{x_1, \ldots, x_n\}$ són una mostra aleatòria d'una distribució normal $\mathrm{N}(x \mid \mu, \sigma)$, on σ se suposa coneguda, i considerem el problema «canònic» de contrastar quan aquestes dades són o no compatibles amb una hipòtesi específica puntual $H_0 \equiv \{\mu = \mu_0\}$ sobre el valor de la mitjana.

L'enfocament convencional d'aquest problema demana una distribució inicial no regular que doni una massa de probabilitat, posem p_0, en el valor μ_0 que volem contrastar, amb la resta de probabilitat $1 - p_0$ repartida de manera contínua sobre $\Re$. Si s'escull que aquesta distribució inicial sigui $p(\mu \mid \mu \neq \mu_0) = N(\mu \mid \mu_0, \sigma_0)$, podem utilitzar el teorema de Bayes per a obtenir la corresponent probabilitat final,

$$\Pr[\mu_0 \mid D, \lambda] = \frac{B_{01}(D, \lambda)\, p_0}{(1 - p_0) + p_0\, B_{01}(D, \lambda)} \ , \tag{42}$$

$$B_{01}(D, \lambda) = \left(1 + \frac{n}{\lambda}\right)^{1/2} \exp\left[-\frac{1}{2}\frac{n}{n + \lambda}\, z^2\right] , \tag{43}$$

on $z = (\overline{x} - \mu_0)/(\sigma/\sqrt{n})$ que mesura, en desviacions estàndard, la distància entre $\overline{x}$ i μ_0, on $\lambda = \sigma^2/\sigma_0^2$ és la raó entre la variància del model i la variància inicial. La funció $B_{01}(D, \lambda)$, una raó de funcions de versemblança (integrades), s'anomena el *factor de Bayes* en favor d'H_0. Amb una funció pèrdua convencional zero-u, H_0 ha de ser rebutjada si $\Pr[\mu_0 \mid D, \lambda] < 1/2$. Per a la seva utilització rutinària, s'ha suggerit a la bibliografia l'elecció de $p_0 = 1/2$ i $\lambda = 1$ o $\lambda = 1/2$ per a descriure formes particulars de coneixement inicial *precís*. L'enfocament convencional del test d'hipòtesis tracta amb situacions de probabilitat inicial *concentrada*; *assumeix* un important coneixement inicial sobre el valor de μ i, d'aquí, *no* pot ser utilitzat llevat que aquesta sigui una assumpció escaient. A més, tal com va indicar Barlett els anys cinquanta, la probabilitat final que resulta és extremadament sensible a l'especificació inicial realitzada. En moltes aplicacions, H_0 es defineix en realitat de forma confusa com una petita regió en comptes d'un punt. Per a grandàries mostrals no molt grans, la probabilitat final $\Pr[\mu_0 \mid D, \lambda]$ és una *aproximació* a la

probabilitat final $\Pr[\mu_0 - \epsilon < \mu < \mu_0 - \epsilon \mid D, \lambda]$ per a algun interval petit al voltant de μ_0 que s'obtindria a partir de distribucions inicials contínues, regulars, fortament concentrades al voltant de μ_0; però aquesta aproximació *mai* funciona per a grandàries mostrals suficientment grans. Una conseqüència (que és immediatament clara a partir de les dues darreres equacions) és que per a qualsevol valor *fixat* de l'estadístic z, la probabilitat final de la hipòtesi nulla, $\Pr[\mu_0 \mid D, \lambda]$, tendeix a 1 quan $n \to \infty$. Lluny de ser específic per a aquest exemple, aquest comportament indesitjable de les probabilitats finals basades en distribucions inicials no regulars puntuals (descobert per Lindley els anys cinquanta, i generalment conegut com a *paradoxa de Lindley*) està *sempre* present en l'enfocament convencional bayesià del test d'hipòtesis *puntuals*.

L'enfocament intrínsec pot ser usat sense suposar cap coneixement inicial puntual. La discrepància intrínseca és $d(\mu_0, \mu) = n(\mu - \mu_0)^2/(2\sigma^2)$, una simple transformació de la distància estandarditzada entre μ i μ_0. Com després veurem (secció 5), la manca d'informació inicial sobre el valor de μ pot formalment ser descrit en aquest problema amb la funció inicial impròpia (uniforme) $p(\mu) = 1$; podem usar el teorema de Bayes per a obtenir la distribució final (pròpia) corresponent, $p(\mu \mid D) = N(\mu \mid \overline{x}, \sigma/\sqrt{n})$. El valor esperat de $d(\mu_0, \mu)$ respecte aquesta distribució final és $\overline{d}(\mu_0, D) = (1 + z^2)/2$, on $z = (\overline{x} - \mu_0)/(\sigma/\sqrt{n})$ és la distància estandaritzada entre $\overline{x}$ i μ_0. Tal com pronosticava la teoria general, el valor esperat de $\overline{d}(\mu_0, D)$ sota mostreig repetit és u si $\mu = \mu_0$, i augmenta linealment amb n si $\mu = \mu_0$. A més, en aquest exemple canònic, rebutjar H_0 quan $|z| > 2$ o $|z| > 3$, que és quan μ_0 és 2 o 3 desviacions estàndards finals lluny de $\overline{x}$, respectivament correspon a rebutjar H_0 quan $\overline{d}(\mu_0, D)$ és més gran que 2,5, o més gran que 5. Però l'escala d'informació és independent del problema, i per tant, rebutjar la hipòtesi nul·la quan la discrepància esperada del veritable model és més gran que $d^* = 5$ unitats de la informació és una regla *general* (que correspon a la regla convencional «3σ» del cas normal canònic).

Si σ és desconeguda, la discrepància intrínseca esdevé

$$d(\mu_0, \mu, \sigma) = \frac{n}{2} \log \left[1 + \left(\frac{\mu - \mu_0}{\sigma} \right)^2 \right]. \tag{44}$$

A més, tal com dèiem abans, la falta d'informació inicial sobre ambdues μ i σ pot ser descrita amb la funció inicial (impròpia) $p(\mu, \sigma) = \sigma^{-1}$. El contrast estadístic intrínsec $\overline{d}(\mu_0, D)$ resulta ser el valor esperat de $d(\mu_0, \mu, \sigma)$ sota la corresponent distribució conjunta final; això es pot expressar amb precisió en termes de funcions hipergeomètriques, i és aproximat per

$$\overline{d}(\mu_0, D) \approx \frac{1}{2} + \frac{n}{2} \log \left(1 + \frac{t^2}{n} \right), \tag{45}$$

on t és l'estadístic tradicional $t = \sqrt{n-1}(\overline{x} - \mu_0)/s$, $ns^2 = \sum_j (x_j - \overline{x})^2$. Per exemple, per a grandàries mostrals 5, 30 i 1.000, i utilitzant la constant d'utilitat $d^* = 5$, la hipòtesi H_0 serà rebutjada quan $|t|$ és respectivament més gran que 5,025, 3,240 i 3,007.

5 Anàlisi de referència

Sota el paradigma bayesià, el resultat de qualsevol problema d'inferència (la distribució final de la quantitat d'interès) combina la informació proporcionada per les dades amb la informació rellevant inicial disponible. En moltes situacions, però, o bé la informació inicial disponible sobre la quantitat d'interès és massa ambigua per a justificar l'esforç demanat per a formalitzar-la com una distribució de probabilitat, o és massa subjectiva per a ser útil en comunicació científica o en presa pública de decisions. Llavors, és important ser capaços d'identificar la forma matemàtica d'una distribució inicial «no informativa», una distribució inicial que tindrà un mínim efecte, relatiu a les dades, sobre la inferència final. Més formalment, suposem que s'assumeix que el mecanisme probabilístic que ha generat les dades disponibles D és $p(D \mid \boldsymbol{\omega})$, per a algun $\boldsymbol{\omega} \in \Omega$, i que la quantitat d'interès és una funció a valors reals $\theta = \theta(\boldsymbol{\omega})$ del paràmetre del model $\boldsymbol{\omega}$. Sense pèrdua de generalitat, podem suposar que el model de probabilitat és de la forma $p(D \mid \theta, \boldsymbol{\lambda})$, $\theta \in \Theta$, $\boldsymbol{\lambda} \in \Lambda$, on $\boldsymbol{\lambda}$ és un vector de paràmetres marginals escollit de manera adient. Tal com hem vist a la secció 3, per a obtenir la distribució final de la quantitat d'interès $p(\theta \mid D)$ és necessari especificar una distribució inicial *conjunta* $p(\theta, \boldsymbol{\lambda})$. Per tant, cal identificar la forma d'aquesta distribució inicial conjunta $\pi_\theta(\theta, \boldsymbol{\lambda})$, la *$\theta$-distribució inicial de referència*, que tindrà un efecte *minimal* en la corresponent distribució final de θ,

$$\pi(\theta \mid D) \propto \int_\Lambda p(D \mid \theta, \boldsymbol{\lambda})\, \pi_\theta(\theta, \boldsymbol{\lambda})\, d\boldsymbol{\lambda}\,, \tag{46}$$

una distribució inicial que, per usar una expressió convencional, «farà que les dades parlin per si mateixes» sobre el valor probable de θ. En rigor, les distribucions *finals* de referència juguen un paper important en la comunicació científica, ja que proporcionen la resposta a una pregunta central en les ciències: condicionat al model assumit $p(D \mid \theta, \boldsymbol{\lambda})$, i amb les altres hipòtesis sobre el valor de θ en les quals hi pugui haver un acord universal, la distribució final de referència $\pi(\theta \mid D)$ ha d'especificar què *podria* ser dit sobre θ *si* l'única informació disponible sobre θ fossin unes dades ben documentades D.

S'ha treballat molt per a formular de manera matemàticament rigorosa la idea anterior de distribucions inicials de «referència». En aquesta secció ens concentrarem en una aproximació basada en la teoria de la informació per a descriure distribucions de referència que es pot justificar que proporcionen el procediment general disponible més avançat. En la formulació descrita més endavant, lluny d'ignorar el coneixement inicial, la distribució final de referència aprofita certes característiques ben definides d'una *possible* distribució inicial, concretament aquelles que descriuen una situació on es pot mantenir que el coneixement rellevant sobre la quantitat d'interès (més enllà del que és universalment acceptat) és negligible comparat amb la informació que és possible proporcionar sobre aquesta quantitat amb l'experimentació repetida (a partir d'un mecanisme particular de generar dades). L'anàlisi de referència

és apropiada en contexts on es considera pertinent el conjunt d'inferències que poden ser obtinguts en aquesta situació.

Qualsevol anàlisi estadística conté un cert nombre d'elements subjectius; entre d'altres, assenyalem les dades seleccionades, les hipòtesis del model i l'elecció de les quantitats d'interès. Es pot argumentar que l'anàlisi de referència proporciona una solució bayesiana «objectiva» als problemes d'inferència estadística en exactament el mateix sentit que els mètodes estadístics convencionals afirmen ser «objectius»: que les solucions només depenen de les hipòtesis del model i les dades observades. Però tot el tema dels mètodes bayesians objectius està sotmès a la polèmica; els lectors interessats trobaran a la bibliografia algunes referències rellevants.

5.1 Distributions de referència

Un paràmetre. Considerem l'experiment que consisteix en l'observació d'unes dades D, generades per un mecanisme aleatori $p(D \mid \theta)$ que només depèn d'un paràmetre real $\theta \in \Theta$, i sigui $\boldsymbol{t} = \boldsymbol{t}(D) \in T$ *qualsevol* estadístic suficient (que pot ser tot el conjunt de dades D). A la teoria general de la informació de Shannon, la *quantitat d'informació* $I^\theta\{T, p(\theta)\}$ que es pot esperar que serà proporcionada per D, o (equivalentment) per a $\boldsymbol{t}(D)$, sobre el valor de θ és defineix amb

$$\begin{aligned} I^\theta\{T, p(\theta)\} &= \int_T \int_\Theta p(\boldsymbol{t}, \theta) \log \frac{p(\boldsymbol{t}, \theta)}{p(\boldsymbol{t})p(\theta)}\, d\theta d\boldsymbol{t} \\ &= \mathrm{E}_{\boldsymbol{t}} \left[\int_\Theta p(\theta \mid \boldsymbol{t}) \log \frac{p(\theta \mid \boldsymbol{t})}{p(\theta)}\, d\theta \right] \end{aligned} \tag{47}$$

la divergència logarítmica esperada de la distribució inicial a partir de la distribució final. Naturalment, aquest és un *funcional* de la distribució inicial $p(\theta)$: com més gran sigui la informació inicial, menor serà la informació que és esperable que proporcionin les dades. El funcional $I^\theta\{T, p(\theta)\}$ és còncau, no negatiu, i invariant sota transformacions bijectives de θ. Considerem ara la quantitat d'informació $I^\theta\{T^k, p(\theta)\}$ sobre θ que es pot esperar d'un experiment que consisteix en k replicacions condicionalment independents $\{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k\}$ de l'experiment original. Quan $k \to \infty$, aquest experiment proporcionarà qualsevol *informació desconeguda* sobre θ que es podria obtenir possiblement en aquest context; així, quan $k \to \infty$, el funcional $I^\theta\{T^k, p(\theta)\}$ s'aproximarà a la informació *desconeguda* sobre θ associada amb la inicial $p(\theta)$. Intuïtivament, una distribució inicial de θ-«no informativa» és una que *maximitza la informació desconeguda* sobre θ. Formalment, si $\pi_k(\theta)$ denota la densitat inicial que maximitza $I^\theta\{T^k, p(\theta)\}$ en la classe $\mathcal{P}$ de distribucions que són compatibles amb les hipòtesis acceptades sobre el valor de θ (que pot ser la classe de *totes* les distribucions inicials pròpies estrictament positives) llavors la inicial de θ-referència $\pi(\theta)$ és el límit quan $k \to \infty$ (en un sentit que cal precisar) de la successió de distribucions inicials $\{\pi_k(\theta), k = 1, 2, \ldots\}$.

Noteu que aquest procediment límit *no* és una mena d'aproximació asimptòtica, sinó un element essencial de la *definició* de distribució inicial de referència. En particular, aquesta definició implica que les distribucions de referència només depenen del comportament *asimptòtic* del model de probabilitat assumit, una característica que en simplifica molt el càlcul efectiu.

11 EXEMPLE *(Màxima entropia.)* Si θ només pot prendre un nombre *finit* de valors, és a dir, quan l'espai de paràmetres és $\Theta = \{\theta_1, \ldots, \theta_m\}$ i $p(\theta) = \{p_1, \ldots, p_m\}$, amb $p_i = \Pr(\theta = \theta_i)$, llavors pot demostrar-se que la informació desconeguda associada a $\{p_1, \ldots, p_m\}$ és

$$\lim_{k\to\infty} I^{\theta}\{T^k, p(\theta)\} = H(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log(p_i)\,, \tag{48}$$

que és l'*entropia* de la distribució inicial $\{p_1, \ldots, p_m\}$.

Així, en el cas finit, la distribució inicial de referència és la que té *màxima entropia* en la classe $\mathcal{P}$ de distribucions inicials compatibles amb les hipòtesis acceptades. En conseqüència, l'algorisme inicial de referència conté distribucions inicials amb «màxima entropia» com a cas particular obtingut quan l'espai de paràmetres és *finit*, l'*únic* cas on el concepte original d'entropia (en mecànica estadística, com a mesura de la incertesa) és no ambigu i es comporta bé. Si, en particular, $\mathcal{P}$ conté *totes* les distribucions inicials sobre $\{\theta_1, \ldots, \theta_m\}$, llavors la distribució inicial de referència és la distribució inicial uniforme, $\pi(\theta) = \{1/m, \ldots, 1/m\}$.

Formalment, la *funció inicial de referència* $\pi(\theta)$ d'un paràmetre univariant θ es defineix com el límit de la successió de distribucions inicials pròpies $\pi_k(\theta)$ que maximitzen $I^{\theta}\{T^k, p(\theta)\}$ en el sentit precís que, per a qualsevol valor de l'estadístic suficient $\boldsymbol{t} = \boldsymbol{t}(D)$, la *distribució final de referència*, el límit puntual $\pi(\theta \mid \boldsymbol{t})$ de la corresponent successió de distribucions finals $\{\pi_k(\theta \mid \boldsymbol{t})\}$, pot ser obtingut de $\pi(\theta)$ amb ús formal del teorema de Bayes i, per tant, $\pi(\theta \mid \boldsymbol{t}) \propto p(\boldsymbol{t} \mid \theta)\, \pi(\theta)$.

Sovint les *funcions* inicials de referència s'anomenen senzillament *distribucions inicials de referència*, fins i tot sabent que normalment *no* són distribucions de probabilitat. *No* s'han de considerar com expressions de creença, sinó com tècniques per a obtenir distribucions finals (pròpies) que són una forma límit de les distribucions finals que podrien haver estat obtingudes a partir de possibles creences inicials que són relativament no informatives respecte a la quantitat d'interès quan es compara amb la informació que les dades podrien proporcionar.

Si *a*) l'estadístic suficient $\boldsymbol{t} = \boldsymbol{t}(D)$ és un estimador consistent $\widetilde{\theta}$ d'un paràmetre continu θ, i *b*) la classe $\mathcal{P}$ conté *totes* les distribucions inicials estrictament positives, llavors pot demostrar-se que la distribució inicial de referència té una forma simple en termes de qualsevol aproximació *asimptòtica* a la distribució final de θ. Noteu que, per construcció, una aproximació *asimptòtica* a

la distribució final *no* depèn de la distribució inicial. Concretament, si la densitat final $p(\theta \mid D)$ té una aproximació asimptòtica de la forma $p(\theta \mid \widetilde{\theta}, n)$, la distribució inicial de referència és simplement

$$\pi(\theta) \propto p(\theta \mid \widetilde{\theta}, n)\Big|_{\widetilde{\theta}=\theta}. \tag{49}$$

Les distribucions inicials de referència uniparamètriques són *invariants* sota reparametritzacions; així, si $\psi = \psi(\theta)$ és una funció bijectiva a trossos de θ, llavors la distribució inicial de ψ-referència és simplement la transformació de probabilitat adient de la distribució inicial de θ-referència.

12 EXEMPLE *(Distribució inicial de Jeffreys.)* Si θ és univariant i continu, i la distribució final de θ suposat $\{x_1 \ldots, x_n\}$ és asimptòticament normal amb desviació típica $s(\widetilde{\theta})/\sqrt{n}$, llavors, a partir de (49), la funció inicial de referència és $\pi(\theta) \propto s(\theta)^{-1}$. Sota condicions de regularitat (que a la pràctica normalment es compleixen, vegeu la secció 3.3), la distribució final de θ és asimptòticament normal amb variància $n^{-1}F^{-1}(\hat{\theta})$, on $F(\theta)$ és la funció d'informació de Fisher i $\hat{\theta}$ és l'estimador del màxim de versemblança de θ. D'aquí, la funció inicial de referència en aquestes condicions és $\pi(\theta) \propto F(\theta)^{1/2}$, que és coneguda com a *distribució inicial de Jeffreys*. D'on es dedueix que l'algorisme inicial de referència conté les distribucions inicials de Jeffreys com a cas particular obtingut quan el model de probabilitat només depèn de paràmetres univariants continus, hi ha condicions de regularitat per a garantir la normalitat asimptòtica, i no hi ha informació addicional, de manera que la classe de totes les possibles distribucions inicials $\mathcal{P}$ conté totes les distribucions inicials sobre Θ estrictament positives. Aquestes són precisament les condicions sota les quals hi ha un acord general d'usar la distribució inicial de Jeffreys com a distribució inicial «no informativa».

2 EXEMPLE *(Inferència sobre un paràmetre binomial, continuació.)* Siguin les dades $D = \{x_1, \ldots, x_n\}$ una successió de n proves independents de Bernoulli, i per tant, $p(x \mid \theta) = \theta^x(1-\theta)^{1-x}$, $x \in \{0, 1\}$; aquest és un model regular, uniparametric continu, que té funció d'informació de Fisher $F(\theta) = \theta^{-1}(1-\theta)^{-1}$. Així, la distribució inicial de referència $\pi(\theta)$ és proporcional a $\theta^{-1/2}(1-\theta)^{-1/2}$, i, per tant, la distribució inicial de referència és la distribució Beta (pròpia) $\mathrm{Be}(\theta \mid 1/2, 1/2)$. Atès que l'algorisme de referència és invariant sota reparametrització, la distribució inicial de referència de $\phi(\theta) = 2 \arcsin\sqrt{\theta}$ és $\pi(\phi) = \pi(\theta)/|\partial\phi/\partial/\theta| = 1$; així, la distribució inicial de referència és *uniforme en la transformació estabilitzadora de la variància* $\phi(\theta) = 2\arcsin\sqrt{\theta}$, una propietat que generalment és veritat sota condicions de regularitat. En termes del paràmetre original θ, la corresponent distribució final de referència és $\mathrm{Be}(\theta \mid r + 1/2, n - r + 1/2)$, on $r = \sum x_j$ és el nombre de proves positives.

Suposem, per exemple, que a $n = 100$ persones seleccionades a l'atzar se'ls ha fet un test per a una infecció i que tots els test són negatius, de manera

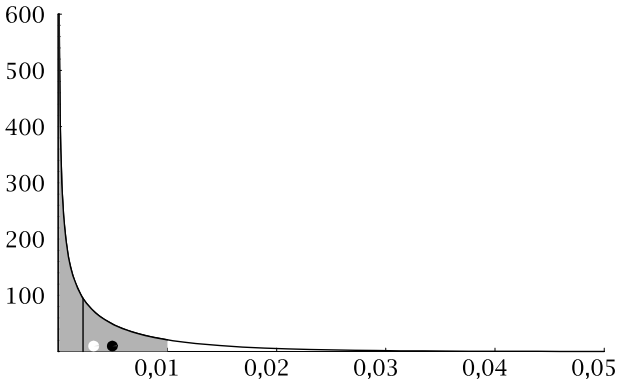

FIGURA 4: Distribució final de la proporció de gent infectada en la població atès que en els resultats de $n = 100$ tests cap d'ells era positiu.

que $r = 0$. La distribució final de referència de la proporció θ de gent infectada és llavors la distribució Beta $\mathrm{Be}(\theta \mid 0{,}5, 100, 5)$, representada a la figura 4. Podria ser conegut que la infecció era estranya, és a dir, suposar que $\theta < \theta_0$, per a alguna fita superior θ_0; la distribució inicial de referència (restringida) serà llavors de la forma $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ si $\theta < \theta_0$, i zero altrament. Però si acceptem que la versemblança està concentrada en la regió $\theta < \theta_0$, llavors la distribució final corresponent serà, de fet, idèntica a $\mathrm{Be}(\theta \mid 0{,}5, 100, 5)$. Així, a partir de la base dels resultats experimentals observats, hom pot afirmar que la proporció de gent infectada és segurament menor que el 5% (ja que la probabilitat final de referència de l'esdeveniment $\theta > 0{,}05$ és 0,001), que θ és menor que 0,01 amb probabilitat 0,844 (àrea de la regió ombrejada a la figura 4), que és igualment probable estar sobre o sota del 0,23% (ja que la mediana, representada per la línia vertical, és 0,0023), i que la probabilitat que una persona escollida a l'atzar de la població estigui infectada és 0,005 (la mitjana final, representada en la figura per un cercle negre), ja que $\Pr(x = 1 \mid r, n) = \mathrm{E}[\theta \mid r, n] = 0{,}005$. Si es necessita una estimació puntual particular de θ (diguem-ne, un nombre per a ser citat a la capçalera de l'informe) l'estimador intrínsec se suggereix a ell mateix, i és $\theta^* = 0{,}0032$ (representat a la figura amb un cercle blanc). Noteu que la solució tradicional a aquest problema, basat en el comportament asimptòtic de l'estimador del màxim de versemblança, aquí $\hat{\theta} = r/n = 0$ per a qualsevol n, és absurd en aquest escenari.

Un paràmetre marginal. L'extensió de l'algorisme inicial al cas de dos paràmetres segueix el procediment matemàtic habitual de reduir el problema a una aplicació seqüencial del procediment establert per al cas uniparamètric.

Així, si el model de probabilitat és $p(\boldsymbol{t} \mid \theta, \lambda)$, $\theta \in \Theta$, $\lambda \in \Lambda$, i volem una distribució inicial de θ-referència $\pi_\theta(\theta, \lambda)$, l'algorisme de referència funciona en dues passes:

a) Condicionat a θ, $p(\boldsymbol{t} \mid \theta, \lambda)$ només depèn del paràmetre marginal λ i, d'aquí, es pot utilitzar l'algorisme uniparamètric per a obtenir la inicial de referència *condicional* $\pi(\lambda \mid \theta)$.

b) Si $\pi(\lambda \mid \theta)$ és pròpia, pot ser usada per a integrar respecte el paràmetre marginal, i se n'obté llavors el model integrat uniparamètric $p(\boldsymbol{t} \mid \theta) = \int_\Lambda p(\boldsymbol{t} \mid \theta, \lambda)\, \pi(\lambda \mid \theta)\, d\lambda$, al qual podem aplicar l'algorisme uniparamètric per a obtenir $\pi(\theta)$. La distribució inicial de θ-referència és llavors $\pi_\theta(\theta, \lambda) = \pi(\lambda \mid \theta)\, \pi(\theta)$, i la distribució final de referència demanada és $\pi(\theta \mid \boldsymbol{t}) \propto p(\boldsymbol{t} \mid \theta)\pi(\theta)$.

Si la distribució inicial de referència condicional *no* és pròpia, llavors el procediment s'executa mitjançant una successió creixent de subconjunts $\{\Lambda_i\}$ que convergeix a Λ, sobre el qual $\pi(\lambda \mid \theta)$ és integrable. Això fa possible obtenir la corresponent successió de distribucions finals de θ-referència $\{\pi_i(\theta \mid \boldsymbol{t}\}$ per a la quantitat d'interès θ, i la distribució final de referència demanada és el corresponent límit puntual $\pi(\theta \mid \boldsymbol{t}) = \lim_i \pi_i(\theta \mid \boldsymbol{t})$. Una distribució inicial de θ-referència és defineix llavors com una funció positiva $\pi_\theta(\theta, \lambda)$ que pot ser utilitzada formalment en el teorema de Bayes com una distribució inicial per a obtenir la distribució final de referència, és a dir, de manera que, per a qualsevol $\boldsymbol{t} \in T$, $\pi(\theta \mid \boldsymbol{t}) \propto \int_\Lambda p(\boldsymbol{t} \mid \theta, \lambda)\, \pi_\theta(\theta, \lambda)\, d\lambda$. Les successions aproximadores han de ser *consistentment* escollides dintre d'un model donat. Així, suposat un model de probabilitat $\{p(\boldsymbol{x} \mid \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$ una successió aproximadora adient $\{\Omega_i\}$ ha de ser escollida per a la totalitat de l'espai de paràmetres Ω; llavors, si l'anàlisi es fa en termes de, posem, $\boldsymbol{\psi} = \{\psi_1, \psi_2\} \in \Psi(\Omega)$, la successió aproximadora ha de ser escollida tal que $\Psi_i = \psi(\Omega_i)$. Una successió aproximadora natural en problemes de posició-escala és $\{\mu, \log \sigma\} \in [-i, i]^2$.

La distribució inicial de θ-referència *no* depèn de l'elecció del paràmetre marginal λ; així, per a qualsevol $\psi = \psi(\theta, \lambda)$ tal que (θ, ψ) és una funció bijectiva de (θ, λ), la distribució inicial de θ-referència en termes de (θ, ψ) és, senzillament, $\pi_\theta(\theta, \psi) = \pi_\theta(\theta, \lambda)/|\partial(\theta, \psi)/\partial(\theta, \lambda)|$, la transformació de probabilitat escaient de la distribució inicial de θ-referència en termes de (θ, λ). Noteu, però, que la distribució inicial de referència *pot* dependre del paràmetre d'interès; així, la distribució inicial de θ-referència pot diferir de la distribució inicial de ϕ-referència llevat que ϕ sigui una transformació bijectiva de θ, o que ϕ sigui asimptòticament independent de θ. Aquesta era una conseqüència esperable del fet que les condicions sota les quals la informació *desconeguda* sobre θ és maximitzada no són generalment les mateixes que les condicions per a maximitzar la informació *desconeguda* sobre alguna funció $\phi = \phi(\theta, \lambda)$.

La *no-existència* d'una única «distribució inicial no informativa» que sigui adient per a qualsevol problema d'inferència en un model donat fou establerta els anys setanta per Dawid i Stone, quan varen provar que era incompatible

amb la *marginalització consistent*. De fet, suposat el model $p(D \mid \theta, \lambda)$, si la distribució final de referència de la quantitat d'interès θ, $\pi(\theta \mid D) = \pi(\theta \mid \boldsymbol{t})$, només depèn de les dades a través d'un estadístic $\boldsymbol{t}$ que té una ditribució mostral, $p(\boldsymbol{t} \mid \theta, \lambda) = p(\boldsymbol{t} \mid \theta)$, que només depèn de θ, es pot esperar que la distribució final de referència sigui de la forma $\pi(\theta \mid \boldsymbol{t}) \propto \pi(\theta)\, p(\boldsymbol{t} \mid \theta)$ per a alguna distribució inicial $\pi(\theta)$. Però es varen trobar exemples on això no podia passar si una *única* distribució inicial «no informativa» conjunta era utilitzada per a qualsevol valor que la quantitat d'interès pogués tenir.

13 EXEMPLE *(Funcions inicial de referència bidimensionals contínues regulars.)* Si la distribució conjunta final de (θ, λ) és asimptòticament normal, llavors la distribució inicial de θ-referència pot ser derivada en termes de la corresponent matriu d'informació de Fisher, $\boldsymbol{F}(\theta, \lambda)$. De fet, si

$$\boldsymbol{F}(\theta, \lambda) = \begin{pmatrix} F_{\theta\theta}(\theta, \lambda) & F_{\theta\lambda}(\theta, \lambda) \\ F_{\theta\lambda}(\theta, \lambda) & F_{\lambda\lambda}(\theta, \lambda) \end{pmatrix}, \quad \text{i} \quad \boldsymbol{S}(\theta, \lambda) = \boldsymbol{F}^{-1}(\theta, \lambda), \tag{50}$$

llavors la distribució inicial de θ-referència és $\pi_\theta(\theta, \lambda) = \pi(\lambda \mid \theta)\, \pi(\theta)$, on

$$\pi(\lambda \mid \theta) \propto F_{\lambda\lambda}^{1/2}(\theta, \lambda), \quad \lambda \in \Lambda. \tag{51}$$

Si $\pi(\lambda \mid \theta)$ és pròpia,

$$\pi(\theta) \propto \exp\{ \int_\Lambda \pi(\lambda \mid \theta)\, \log[S_{\theta\theta}^{-1/2}(\theta, \lambda)]\, d\lambda \}, \quad \theta \in \Theta. \tag{52}$$

Si $\pi(\lambda \mid \theta)$ no és pròpia, llavors, a partir d'integracions sobre una successió aproximadora $\{\Lambda_i\}$, s'obté una successió $\{\pi_i(\lambda \mid \theta)\, \pi_i(\theta)\}$, (on $\pi_i(\lambda \mid \theta)$ és la renormalització pròpia de $\pi(\lambda \mid \theta)$ en Λ_i) i la distribució inicial de θ-referència $\pi_\theta(\theta, \lambda)$ es defineix com el límit adient. A més, si *a*) ambdues $F_{\lambda\lambda}^{1/2}(\theta, \lambda)$ i $S_{\theta\theta}^{-1/2}(\theta, \lambda)$ *factoritzen*, això és,

$$S_{\theta\theta}^{-1/2}(\theta, \lambda) \propto f_\theta(\theta)\, g_\theta(\lambda), \quad F_{\lambda\lambda}^{1/2}(\theta, \lambda) \propto f_\lambda(\theta)\, g_\lambda(\lambda), \tag{53}$$

i *b*) els paràmetres θ i λ són *de variació independent*, és a dir, si Λ no depèn de θ, llavors la distribució inicial de θ-referència és senzillament $\pi_\theta(\theta, \lambda) = f_\theta(\theta) g_\lambda(\lambda)$, fins i tot si la distribució inicial de referència condicional $\pi(\lambda|\theta) = \pi(\lambda) \propto g_\lambda(\lambda)$ (que no depèn de θ) és, de fet, impròpia.

3 EXEMPLE *(Inferència sobre paràmetres normals, continuació.)* La matriu d'informació que correspon al model normal $\mathrm{N}(x \mid \mu, \sigma)$ és

$$\boldsymbol{F}(\mu, \sigma) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{pmatrix}, \quad \boldsymbol{S}(\mu, \sigma) = \boldsymbol{F}^{-1}(\mu, \sigma) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{1}{2}\sigma^2 \end{pmatrix}; \tag{54}$$

d'aquí $F_{\sigma\sigma}^{1/2}(\mu, \sigma) = \sqrt{2}\, \sigma^{-1} = f_\sigma(\mu)\, g_\sigma(\sigma)$, amb $g_\sigma(\sigma) = \sigma^{-1}$, i així $\pi(\sigma \mid \mu) = \sigma^{-1}$. Anàlogament, $S_{\mu\mu}^{-1/2}(\mu, \sigma) = \sigma^{-1} = f_\mu(\mu)\, g_\mu(\sigma)$, amb $f_\mu(\mu) = 1$, i així $\pi(\mu) = 1$. Llavors, la distribució inicial de μ-referència és

$$\pi_\mu(\mu, \sigma) = \pi(\sigma \mid \mu)\pi(\mu) = \sigma^{-1},$$

com ja havíem anticipat. A més, tal com podríem esperar del fet que $\boldsymbol{F}(\mu,\sigma)$ és diagonal, i també anticipat, de manera similar es troba que la distribució inicial de σ-referència és $\pi_\sigma(\mu,\sigma)=\sigma^{-1}$, la mateixa que $\pi_\mu(\mu,\sigma)$.

Suposem, però, que la quantitat d'interès *no* és la mitjana μ o la desviació típica σ, sinó la mitjana *estandarditzada* $\phi=\mu/\sigma$. La matriu d'informació de Fisher en termes dels paràmetres ϕ i σ és $\boldsymbol{F}(\phi,\sigma)=J^t\,\boldsymbol{F}(\mu,\sigma)\,J$, on $J=(\partial(\mu,\sigma)/\partial(\phi,\sigma))$ és el jacobià de la transformació inversa; i llavors dóna

$$\begin{aligned}\boldsymbol{F}(\phi,\sigma)&=\begin{pmatrix}1 & \phi\sigma^{-1}\\ \phi\sigma^{-1} & \sigma^{-2}(2+\phi^2)\end{pmatrix},\\ \boldsymbol{S}(\phi,\sigma)&=\begin{pmatrix}1+\frac{1}{2}\phi^2 & -\frac{1}{2}\phi\sigma\\ -\frac{1}{2}\phi\sigma & \frac{1}{2}\sigma^{-2}\end{pmatrix}.\end{aligned}\tag{55}$$

Així, $S_{\phi\phi}^{-1/2}(\phi,\sigma)\propto(1+\frac{1}{2}\phi^2)^{-1/2}$ i $F_{\sigma\sigma}^{1/2}(\phi,\sigma)\propto\sigma^{-1}(2+\phi^2)^{1/2}$. D'aquí, utilitzant un altre cop els resultats de l'exemple 13, $\pi_\phi(\phi,\sigma)=(1+\frac{1}{2}\phi^2)^{-1/2}\sigma^{-1}$. En la parametrització original, això és $\pi_\phi(\mu,\sigma)=(1+\frac{1}{2}(\mu/\sigma)^2)^{-1/2}\sigma^{-2}$, que és diferent de $\pi_\mu(\mu,\sigma)=\pi_\sigma(\mu,\sigma)$. La corresponent distribució final de ϕ és $\pi(\phi\mid x_1,\ldots,x_n)\propto(1+\frac{1}{2}\phi^2)^{-1/2}\,p(t\mid\phi)$ on $t=(\sum x_j)/(\sum x_j^2)^{1/2}$, un estadístic unidimensional que té distribució mostral, $p(t\mid\mu,\sigma)=p(t\mid\phi)$, que només depèn de ϕ. Per tant, l'algorisme inicial de referència és consistent sota marginalització.

Diversos paràmetres. L'algorisme de referència es generalitza fàcilment a un nombre arbitrari de paràmetres. Si el model és $p(\boldsymbol{t}\mid\omega_1,\ldots,\omega_m)$, es pot obtenir seqüèncialment una distribució inicial de referència

$$\pi(\theta_m\mid\theta_{m-1},\ldots,\theta_1)\times\ldots\times\pi(\theta_2\mid\theta_1)\times\pi(\theta_1)\tag{56}$$

per a cada parametrització *ordenada* $\{\theta_1(\boldsymbol{\omega}),\ldots,\theta_m(\boldsymbol{\omega})\}$ d'interès, i aquestes són invariants per reparametrització de qualsevol dels $\theta_i(\boldsymbol{\omega})$. L'elecció de la parametrització ordenada $\{\theta_1,\ldots,\theta_m\}$ precisament descriu la inicial particular buscada, aquella que *seqüencialment* maximitza la informació desconeguda sobre cada un dels θ_i, condicionats a $\{\theta_1,\ldots,\theta_{i-1}\}$, per a $i=m,\ m-1,\ldots,1$.

14 EXEMPLE *(Paradoxa d'Stein.)* Sigui D una mostra aleatòria d'una distribució normal m-variada amb mitjana $\boldsymbol{\mu}=\{\mu_1,\ldots,\mu_m\}$ i matriu de variàncies unitària. La inicial de referència que correspon a qualsevol permutació de les μ_i és uniforme, i aquesta inicial proporciona, de fet, les distribucions de referència finals adients per a cadascun dels μ_i, que són $\pi(\mu_i\mid D)=N(\mu_i\mid\overline{x}_i,1/\sqrt{n})$. Suposem, però, que la quantitat d'interès és $\theta=\sum_i\mu_i^2$, la norma de $\boldsymbol{\mu}$ a l'origen. Tal com va ser provat per Stein els anys cinquanta, la distribució final de θ basada en aquesta distribució inicial uniforme (o en qualsevol aproximació impròpia «plana») té propietats molt indesitjables; això es deu al fet

que una distribució inicial uniforme (o quasi uniforme), malgrat que és «no informativa» respecte de cadascun dels μ_i individuals, és en veritat altament informativa sobre la suma dels seus quadrats, i, per tant, introdueix un fort biaix positiu (paradoxa d'Stein). Però la distribució inicial de referència que correspon a una reparametrització de la forma $\{\theta, \lambda_1, \ldots, \lambda_{m-1}\}$ produeix, per a qualsevol elecció dels paràmetres marginals $\lambda_i = \lambda_i(\boldsymbol{\mu})$, la distribució final de referència $\pi(\theta \mid D) = \pi(\theta \mid t) \propto \theta^{-1/2}\chi^2(nt \mid m, n\theta)$, on $t = \sum_i \overline{x}_i^2$, i es comprova que aquesta distribució final té propietats de consistència adients.

El mal comportament de distribucions finals marginals concretes derivades de distribucions inicials multivariants «planes» (pròpies o impròpies) en problemes amb molts paràmetres no és específic de l'exemple d'Stein, sinó que és molt freqüent. En conseqüència, es desaconsella fortament la utilització poc crítica o mandrosa de distribucions inicials «planes» en lloc de distribucions inicials de referència rellevants.

Informació limitada. Malgrat la seva utilització en contexts on no hi ha acord universal sobre el coneixement inicial que hi ha disponible sobre la quantitat d'interès, es pot utilitzar l'algorisme de referència per a especificar una distribució inicial que incorpori tot el coneixement inicial acceptable; només cal maximitzar la informació desconeguda dintre de la classe $\mathcal{P}$ de distribucions inicials que són compatibles amb aquest coneixement acceptat. De fet, amb la incorporació progressiva de restriccions addicionals en $\mathcal{P}$, l'algorisme inicial de referència esdevé un mètode d'*assignació de probabilitat* inicial). Tal com hem descrit abans, el problema té una solució analítica simple força bona quan aquelles restriccions tenen la forma de valors esperats coneguts. La incorporació d'altres classes de restriccions normalment implica càlculs numèrics.

15 EXEMPLE *(Distribucions inicials de referència restringides univariants.)* Si suposem que el mecanisme probabilístic que ha generat les dades disponibles només depèn d'un paràmetre univariant continu $\theta \in \Theta \subset \Re$, i la classe $\mathcal{P}$ de distribucions inicials acceptables és una classe de distribucions inicials pròpies que compleix unes restriccions esperables en els seus valors, de manera que

$$\begin{aligned} \mathcal{P} = \Big\{ & p(\theta)\,; \quad p(\theta) > 0, \int_\Theta p(\theta)\, d\theta = 1\,, \\ & \int_\Theta g_i(\theta)\, p(\theta)\, d\theta = \beta_i,\ i = 1, \ldots, m \Big\}\,, \end{aligned} \tag{57}$$

llavors la distribució inicial de referència (restringida) és

$$\pi(\theta \mid \mathcal{P}) \propto \pi(\theta)\, \exp\left[\sum_{j=1}^{m} \gamma_i\, g_i(\theta)\right] \tag{58}$$

on $\pi(\theta)$ és la distribució inicial de referència no restringida i les γ_i són constants (els multiplicadors de Lagrange corresponents), que seran determinats

per les restriccions que defineixen $\mathcal{P}$. Suposem, per exemple, que considerem les dades com una mostra aleatòria d'un model centrat en θ i que, a més, suposem que $\mathrm{E}[\theta] = \mu_0$ i que $\mathrm{Var}[\theta] = \sigma_0^2$. Pot demostrar-se que la distribució inicial de referència no restringida per a qualsevol problema de posició regular és la uniforme. Així, la distribució inicial de referència no restringida ha de ser de la forma $\pi(\theta \mid \mathcal{P}) \propto \exp\{\gamma_1 \theta + \gamma_2(\theta - \mu_0)^2\}$, amb $\int_\Theta \theta\, \pi(\theta \mid \mathcal{P})\, d\theta = \mu_0$ i $\int_\Theta (\theta - \mu_0)^2\, \pi(\theta \mid \mathcal{P})\, d\theta = \sigma_0^2$. D'aquí, $\pi(\theta \mid \mathcal{P})$ és una distribució *normal* amb la mitjana i variància especificada.

5.2 Propietats freqüentistes

Els mètodes bayesians proporcionen una solució *directa* a problemes típics d'inferència estadística; de fet, les distribucions finals precisament exposen allò que pot ser dit sobre les quantitats desconegudes d'interès amb les dades disponibles i el coneixement inicial adient. En particular, les distribucions finals de referència no restringides estableixen allò que pot ser dit sense coneixement inicial sobre la quantitat d'interès.

Malgrat tot, pot ser il·luminadora una anàlisi freqüentista del comportament dels procediments bayesians sota mostreig repetit, ja que proporciona alguns ponts interessants entre la inferència freqüentista i la bayesiana. Llavors es troba que les propietats freqüentistes dels procediments de referència bayesians són normalment excel·lents, i poden ser usats per a proporcionar una forma de calibració de probabilitats finals de referència.

Estimació puntual. Generalment s'accepta que, quan la grandària mostral augmenta, un «bon» estimador $\tilde{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ ha de tendir al valor correcte de $\boldsymbol{\theta}$, és a dir, que és *consistent*. Sota condicions de regularitat adients, qualsevol estimador de Bayes $\boldsymbol{\phi}^*$ d'una funció $\boldsymbol{\phi}(\boldsymbol{\theta})$ convergeix en probabilitat a $\boldsymbol{\phi}(\boldsymbol{\theta})$, de manera que les successions d'estimadors de Bayes són típicament consistents. De fet, és conegut que si existeix una successió consistent d'estimadors, llavors els estimadors de Bayes són consistents. La velocitat de convergència sovint és millor per als estimadors de Bayes de referència.

També s'accepta generalment que un «bon» estimador ha de ser *admissible*, és a dir, *no dominat* per cap altre estimador, cosa que vol dir que la pèrdua esperada sota mostreig (condicionat a $\boldsymbol{\theta}$) no pot ser més gran per a tots els valors de $\boldsymbol{\theta}$ que la corresponent a un altre estimador. Tot estimador de Bayes *propi* és admissible; a més, tal com va demostrar Wald els anys cinquanta, un procediment *ha de ser* bayesià (propi o impropi) per a ser admissible. Molts dels resultats publicats sobre admissibilitat es refereixen a funcions de pèrdua quadràtica, però sovint s'estenen a funcions de pèrdua més generals. Els estimadors de Bayes de referència són normalment admissibles respecte de funcions de pèrdua intrínseca.

Cal remarcar que s'ha comprovat que moltes altres idees freqüentistes aparentment intuïtives en estimació eren potencialment enganyoses. Per exemple, si donada una successió de n observacions de Bernoulli amb paràme-

tre θ s'observen r proves positives, la *millor* estimació sense biaix de θ^2 és $r(r-1)/\{n(n-1)\}$, que condueix a $\widetilde{\theta}^2 = 0$ quan $r = 1$; però estimar com a zero la probabilitat de dues proves positives, quan s'ha observat una prova positiva, no és gaire raonable. En fort contrast, qualsevol estimador de Bayes de referència proporciona una resposta raonable. Per exemple, l'estimador intrínsec de θ^2 és simplement $(\theta^*)^2$, on θ^* és l'estimador intrínsec de θ descrit a la secció 4.1. En particular, si $r = 1$ i $n = 2$ l'estimador intrínsec de θ^2 és (tal com es podria esperar) $(\theta^*)^2 = 1/4$.

Estimació per interval. Quan la grandària mostral augmenta, la probabilitat freqüentista de cobriment d'una regió final q-creïble normalment convergeix a q, de manera que, per a *mostres grans*, els intervals bayesians creïbles poden (sota condicions de regularitat) ser interpretats com a regions de confiança freqüentistes *aproximades*: sota mostreig repetit, una regió bayesiana q-creïble de $\boldsymbol{\theta}$ basada en una mostra gran cobrirà el valor veritable de $\boldsymbol{\theta}$ aproximadament $100q\,\%$ de vegades. S'obtenen ràpidament resultats detallats per a problemes univariants. Per exemple, considerem el model de probabilitat $\{p(D \mid \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, sigui $\theta = \theta(\boldsymbol{\omega})$ una quantitat univariant d'interès, i sigui $\boldsymbol{t} = \boldsymbol{t}(D) \in T$ un estadístic suficient. Si $\theta_q(\boldsymbol{t})$ denota el $100q\,\%$ quantil de la distribució final de θ que correspon a alguna inicial no especificada, de manera que

$$\Pr[\theta \le \theta_q(\boldsymbol{t}) \mid \boldsymbol{t}] = \int_{\theta \le \theta_q(\boldsymbol{t})} p(\theta \mid \boldsymbol{t})\, d\theta = q\,, \tag{59}$$

llavors la probabilitat de cobriment de l'interval q-creïble $\{\theta; \theta \le \theta_q(\boldsymbol{t})\}$,

$$\Pr[\theta_q(\boldsymbol{t}) \ge \theta \mid \boldsymbol{\omega}] = \int_{\theta_q(\boldsymbol{t}) \ge \theta} p(\boldsymbol{t} \mid \boldsymbol{\omega})\, d\boldsymbol{t}\,, \tag{60}$$

normalment compleix que $\Pr[\theta_q(\boldsymbol{t}) \ge \theta \mid \boldsymbol{\omega}] = \Pr[\theta \le \theta_q(\boldsymbol{t}) \mid \boldsymbol{t}] + O(n^{-1/2})$. Aquesta aproximació *asimptòtica* és veritat per a *qualsevol* distribució inicial positiva (suficientment regular). Però l'aproximació és millor, de fet és $O(n^{-1})$, per a classes particulars de distribucions inicials conegudes com a *distribucions inicials amb probabilitats aparellades* (primer ordre). Les distribucions inicials de referència són normalment de probabilitats aparellades, de manera que proporcionen aquest acord asimptòtic millorat. De fet, l'acord (en problemes regulars) és normalment força bo, fins i tot en grandàries mostrals moderades.

16 EXEMPLE *(Producte de mitjanes normals.)* Considerem el cas que es prenen mostres aleatòries independents $\{x_1, \ldots, x_n\}$ i $\{y_1, \ldots, y_m\}$ de densitats normals $N(x \mid \omega_1, 1)$ i $N(y \mid \omega_2, 1)$, i suposem que la quantitat d'interès és el producte de les seves mitjanes, $\phi = \omega_1\omega_2$ (per exemple, es volen fer inferències sobre l'àrea ϕ d'un terreny rectangular a partir de les mesures $\{x_i\}$ i $\{y_j\}$ dels costats). Noteu que es tracta d'una versió simplificada del problema habitual en ciència quan s'està interessat en el producte de diverses magnituds, totes

mesurades amb error. A partir del procediment descrit a l'exemple 13, amb la successió aproximadora natural induïda per $(\omega_1, \omega_2) \in [-i, i]^2$, la distribució inicial de ϕ-referència és

$$\pi_\phi(\omega_1, \omega_2) \propto (n\,\omega_1^2 + m\,\omega_2^2)^{-1/2}, \tag{61}$$

molt diferent de la distribució inicial uniforme $\pi_{\omega_1}(\omega_1, \omega_2) = \pi_{\omega_2}(\omega_1, \omega_2) = 1$ que hauria de ser usada per a fer inferències objectives sobre ω_1 o ω_2. Pot demostrar-se que la distribució inicial $\pi_\phi(\omega_1, \omega_2)$ proporciona un acord aproximat entre regions creïbles bayesianes i intervals de confiança freqüentistes per a ϕ; de fet, aquesta distribució inicial va ser suggerida inicialment per Stein els anys vuitanta precisament per a obtenir aquest acord aproximat. El mateix exemple va ser finalment usat per Efron per a reforçar el fet que, fins i tot en un model de probabilitat fixat, $\{p(D \mid \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, la distribució inicial demanada per a fer inferències objectives sobre una funció dels paràmetres $\phi = \phi(\boldsymbol{\omega})$ normalment ha de dependre de la funció ϕ.

L'acord numèric entre regions bayesianes de referència creïbles i intervals de confiança freqüentistes és de fet perfecte en circumstàncies especials. Tal com Lindley va assenyalar els anys cinquanta, aquest és el cas en aquells problemes d'inferència que poden ser transformats en problemes de posició i escala.

3 EXEMPLE *(Inferència sobre paràmetres normals, continuació.)* Sigui $D = \{x_1, \ldots, x_n\}$ una mostra aleatòria d'una distribució normal $N(x \mid \mu, \sigma)$. Tal com dèiem abans, la distribució final de referència de la quantitat d'interès μ és la distribució d'Student $\mathrm{St}(\mu \mid \overline{x}, s/\sqrt{n-1}, n-1)$. Llavors, normalitzant μ, la distribució *final* de $t(\mu) = \sqrt{n-1}(\overline{x} - \mu)/s$, com una funció de μ donat D, és la distribució d'Student estàndard $\mathrm{St}(t \mid 0, 1, n-1)$ amb $n-1$ graus de llibertat. D'altra banda, aquesta funció t és precisament l'estadístic convencional t, que té per *distribució mostral* també la distribució d'Student amb $n-1$ graus de llibertat. Es dedueix que, *per a qualsevol grandària mostral*, els intervals creïbles de *referència* final per a μ donades les dades seran *numèricament idèntics* als intervals de confiança freqüentistes basats en la distribució mostral de t.

Un resultat similar s'obté en inferències sobre la variància. Així, la distribució de referència final de $\lambda = \sigma^{-2}$ és Gamma $\mathrm{Ga}(\sigma^{-2} \mid (n-1)/2, ns^2/2)$ i, d'aquí, la distribució final de $r = ns^2/\sigma^2$, com una funció de σ^2 donat D, és una χ^2 (central) amb $n-1$ graus de llibertat. Però la funció r és l'estadístic convencional per a aquest problema, que té *distribució mostral també* una χ^2 amb $n-1$ graus de llibertat. D'on, *per a qualsevol grandària mostral*, els intervals finals de *referència* creïbles per a σ^2 (o qualsevol funció bijectiva de σ^2) ateses les dades seran *numèricament idèntics* als intervals de confiança freqüentistes basats en la distribució mostral de r.

6 Estudi d'un cas simplificat

Per a il·lustrar els aspectes més importants dels mètodes bayesians, i per a proporcionar un exemple complet i detallat, analitzarem una versió simplificada d'un problema d'enginyeria.

Per a estudiar la fiabilitat d'un gran lot de producció, es fa una prova cara i destructiva a n ítems escollits a l'atzar, i s'obtenen llurs temps de vida observats en hores d'ús continu, $D = \{x_1, \ldots, x_n\}$. Diferents consideracions suggereixen que podríem suposar que el temps de vida x_i de cada ítem segueix una llei exponencial amb taxa de risc θ, és a dir, $p(x_i \mid \theta) = \mathrm{Ex}[x_i \mid \theta] = \theta e^{-\theta x_i}$, $\theta > 0$, i que, donat θ, els temps de vida dels n ítems són independents. Els enginyers de qualitat estan interessats en la informació sobre el valor real de la taxa de risc θ, i en la predicció del temps de vida x d'ítems similars. En particular, volen conèixer la compatibilitat de les dades observades amb els valors anunciats de la taxa de risc i en la proporció d'ítems que podem esperar que tinguin una vida més gran que una especificació industrial demanada.

L'anàlisi estadística de dades exponencials utilitza la distribució exponencial-gamma $\mathrm{Eg}(x \mid \alpha, \beta)$, obtinguda com una mixtura contínua d'exponencials amb una densitat gamma,

$$\mathrm{Eg}(x \mid \alpha, \beta) = \int_0^\infty \theta e^{-\theta x}\, \mathrm{Ga}(\theta \mid \alpha, \beta)\, d\theta = \frac{\alpha \beta^\alpha}{(x+\beta)^{\alpha+1}}\,,$$

$$x \geq 0, \quad \alpha > 0, \quad \beta > 0\,. \quad (62)$$

Aquesta és una densitat monòtona decreixent amb moda al zero; si $\alpha > 1$, té mitjana $\mathrm{E}[x \mid \alpha, \beta] = \beta/(\alpha - 1)$. A més, les probabilitats de les cues tenen una expressió simple; concretament,

$$\Pr[x > t \mid \alpha, \beta] = \left\{ \frac{\beta}{\beta + t} \right\}^\alpha . \quad (63)$$

Funció de versemblança. Sota les hipòtesis acceptades sobre el mecanisme que ha generat les dades, $p(D \mid \theta) = \prod_j \theta e^{-\theta x_j} = \theta^n e^{-\theta s}$, que només depèn de la suma de les observacions $s = \sum_j x_j$. Així, $\boldsymbol{t} = (s, n)$ és un estadístic *suficient* per a aquest model. El corresponent estimador del màxim de versemblança és $\widehat{\theta} = n/s$ i la funció d'informació de Fisher és $F(\theta) = \theta^{-2}$. A més, la distribució mostral de s és la distribució Gamma $p(s \mid \theta) = \mathrm{Ga}(s \mid n, \theta)$.

Les dades reals consisteixen en $n = 25$ temps de vida observats no censurats, en milers d'hores, que dóna una suma $s = 41.574$, i d'aquí la mitjana és $\overline{x} = 1.663$, i l'estimador del màxim de versemblança és $\widehat{\theta} = 0{,}601$. La desviació típica del temps de vida observat fou 1.286 i el seu rang $[0{,}136, 5{,}591]$, que mostra la gran variabilitat (des de centenars a milers d'hores) típicament observada en dades exponencials.

A partir dels resultats de la secció 3.3 i la forma de la funció d'informació de Fisher que hem vist abans, la distribució final *assimptòtica* de θ és

$p(\theta \mid D) \approx \mathrm{N}(\theta \mid \hat{\theta}, \hat{\theta}/\sqrt{n}) = \mathrm{N}(\theta \mid 0{,}601, 0{,}120)$. D'aquí es pot obtenir una primera aproximació ràpida dels possibles valors de θ que, per exemple, podríem esperar que pertanyessin a l'interval $0{,}601 \pm 1{,}96 \cdot 0{,}120$, o $(0{,}366, 0{,}837)$, amb probabilitat propera a 0,95.

6.1 Anàlisi bayesiana objectiva

S'ha d'auditar una empresa a petició d'un client important. Cal preparar un report amb la informació disponible de la taxa de risc θ a partir *exclusivament* de dades *documentades* D, com si aquestes fossin l'*única* informació disponible. Dintre del context bayesià, aquesta anàlisi «objectiva» (objectiva aquí vol dir que no s'utilitza cap més informació que la que proporcionen les dades sota el model assumit) pot ser assolida calculant la corresponent distribució final de *referència*.

Distribucions inicials i finals de referència. El model exponencial compleix totes les condicions de regularitat necessàries. Així, a partir dels resultats de l'exemple 12 i la forma de la funció d'informació de Fisher citada anteriorment, la *funció inicial de referència* (que en aquest cas també és la distribució inicial de Jeffreys) és simplement $\pi(\theta) \propto F(\theta)^{1/2} = \theta^{-1}$. D'aquí, amb teorema de Bayes, la distribució final de referència és $\pi(\theta \mid D) \propto p(D|\theta)\,\theta^{-1} \propto \theta^{n-1}\, e^{-s\theta}$, el nucli de la densitat gamma, és a dir,

$$\pi(\theta \mid D) = \mathrm{Ga}(\theta \mid n, s), \quad \theta > 0\,, \tag{64}$$

que té mitjana $\mathrm{E}[\theta \mid D] = n/s$ (que és també l'estimador del màxim de versemblança $\hat{\theta}$), moda $(n-1)/s$, i desviació típica $\sqrt{n}/s = \hat{\theta}/\sqrt{n}$. Així, la distribució final de referència de la taxa de risc és $\pi(\theta \mid D) = \mathrm{Ga}(\theta \mid 25, 41, 57)$ (representada a la figura 5) amb mitjana 0,601, moda 0,577, i desviació típica 0,120. Una integració numèrica unidimensional proporciona $\Pr[\theta < 0{,}593 \mid D] = 0{,}5$, $\Pr[\theta < 0{,}389 \mid D] = 0{,}025$ i $\Pr[\theta < 0{,}859 \mid D] = 0{,}975$, la mediana és 0,593, i l'interval $(0{,}389, 0{,}859)$ és una regió final de referència 95% creïble (àrea ombrejada a la figura 5). L'estimador intrínsec (vegeu més endavant) és 0,590 (línia discontínua a la figura 5).

Sota les hipòtesis acceptades per al mecanisme probabilístic que han generat les dades, la distribució final de referència $\pi(\theta \mid D) = \mathrm{Ga}(\theta \mid 25, 41, 57)$ conté *tot* allò que podria ser dit sobre el valor de la taxa de risc θ atenent-nos exclusivament a les dades observades D. La figura 5 i els nombres citats més amunt proporcionen, respectivament, resums gràfics i numèrics; però el consultor estadístic va explicar als enginyers el fet que $\pi(\theta \mid D)$ és la resposta *completa* (necessària per a poder continuar treballant en predicció o presa de decisions).

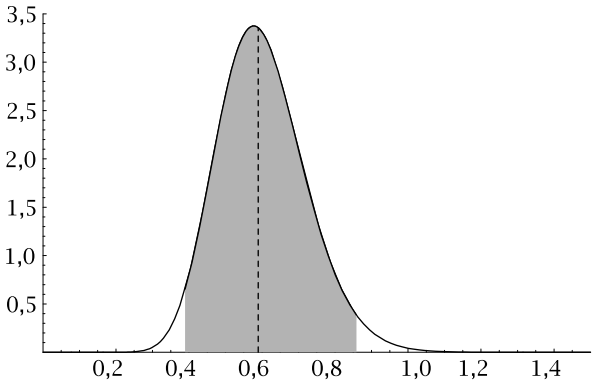

FIGURA 5: Densitat final de referència de la taxa de risc θ. La regió ombrejada és un interval creïble del 95 %. La línia discontínua indica la posició de l'estimador intrínsec.

Distribució predictiva final de referència. La densitat final predictiva de referència del temps de vida x és

$$\pi(x \mid D) = \int_0^\infty \theta e^{-\theta x} \, \mathrm{Ga}(\theta \mid n, s) \, d\theta = \mathrm{Eg}(\theta \mid n, s) \tag{65}$$

amb mitjana $s/(n-1)$. Així, la densitat predictiva final del temps de vida d'un ítem aleatori produït en condicions similars és $\pi(x \mid D) = \mathrm{Eg}(x \mid 25, 41, 57)$, representat a la figura 6 sobre l'histograma de les dades observades. La mitjana d'aquesta distribució és 1,732; d'aquí, donades les dades D, el temps de vida esperat d'ítems similars és 1,732 milers d'hores. El contracte amb el client especificava una compensació per a cada ítem que tingués un temps de vida inferior a 250 hores. Atès que

$$\Pr[x < b \mid D] = \int_0^b \mathrm{Eg}(x \mid n, s) = 1 - \left\{ \frac{s}{s+b} \right\}^n, \tag{66}$$

la proporció d'ítems amb temps de vida menor que 250 és $\Pr[x < 0{,}250 \mid D] = 0{,}139$, l'àrea ombrejada de la figura 6. Així, condicionat a les hipòtesis acceptades, es va dir als enginyers d'esperar un 14 % d'ítems no conformes.

Calibració. Considerem $t = t(\theta) = (s/n)\theta$ com una funció de θ, i la seva transformació inversa $\theta = \theta(t) = (n/s)t$. Atès que $t = t(\theta)$ és una transformació bijectiva de θ, si R_t és una regió final q-creïble per a t, llavors $R_\theta = \theta(R_t)$ és una regió final q-creïble per a θ. A més, fent un canvi de variables, la distribució *final* de referència de $t = t(\theta)$, com una funció de θ condicionada a s, és $\pi(t(\theta) \mid n, s) = \pi(\theta \mid n, s)/|\partial t(\theta)/\partial\theta| = \mathrm{Ga}(t \mid n, n)$,

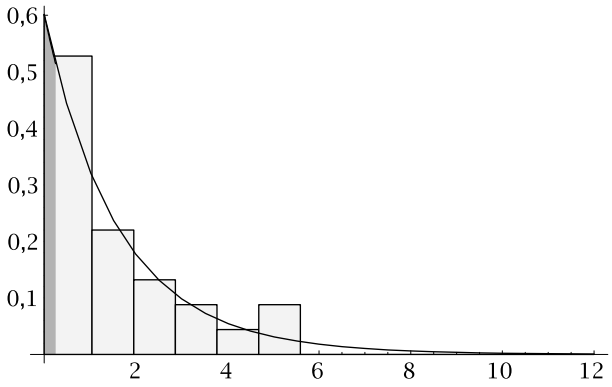

FIGURA 6: Densitat predictiva final de referència del temps de vida (en milers d'hores). La regió ombrejada representa la probabilitat de produir ítems no conformes amb temps de vida menor de 250 hores. El fons és un histograma de les dades observades.

una densitat gamma que no depén de s. D'altra banda, la distribució mostral de l'estadístic suficient s és $p(s \mid n, \theta) = \text{Ga}(\theta \mid n, \theta)$; llavors, la distribució *mostral* de $t = t(s) = (\theta/n)s$, com una funció de s condicionada a θ, és $p(t(s) \mid n, \theta) = p(s \mid n, \theta)/|\partial t(s)/\partial s| = \text{Ga}(t \mid n, n)$, que no conté θ i és precisament la *mateixa* densitat gamma obtinguda abans. D'on, per a *qualsevol* mostra de grandària n, totes les regions finals de referència q-creïbles de la taxa de risc θ seran *també* regions de confiança freqüentista de nivell q. Qualsevol regió final de referència q-creïble té, donades les dades, un grau (racional) de creença q de contenir el valor veritable de θ; el resultat que acabem d'obtenir pot ser utilitzat per a fer una calibració exacta per a aquest grau de creença. De fet, per a qualsevol $\theta > 0$ i qualsevol $q \in (0, 1)$, la proporció límit de regions finals de referència q-creïbles que contenen el valor vertader de θ sota mostreig repetit és precisament igual a q. Llavors es pot explicar als enginyers que quan informin que s'espera que la taxa de risc θ de la seva producció pertanyi a $(0{,}389, 0{,}859)$ amb probabilitat (grau racional de creença) $0{,}95$, podrien afirmar que és una declaració *calibrada*, és a dir, que hipotètiques replicacions del mateix *procediment* sota condicions controlades, amb mostres simulades a partir de *qualsevol* distribució exponencial, donarà un $95\,\%$ de regions que contenen el valor a partir del qual la mostra va estar simulada.

Estimació. El departament comercial podria utilitzar qualsevol mesura de posició de la distribució final de referència de θ com un estimador intuïtiu $\widetilde{\theta}$ de la taxa de risc θ, però si cal escollir un valor particular amb, diguem, alguna importància legal, això implica un problema de decisió per al qual s'ha

d'especificar una funció de pèrdua $L(\widetilde{\theta}, \theta)$. Tot i que no hi havia cap decisió particular prevista, l'empresa auditora va demanar que s'esmentés al report un estimador particular, i per a justificar l'elecció de l'estimador *intrínsec* se'n varen explicar les atractives propietats. La discrepància intrínseca $d(\theta_i, \theta_j)$ entre els *models* $\mathrm{Ex}(x \mid \theta_i)$ i $\mathrm{Ex}(x \mid \theta_j)$ és

$$d(\theta_i, \theta_j) = \min\{\delta(\theta_i \mid \theta_j),\, \delta(\theta_j \mid \theta_i)\}\,, \tag{67}$$

on

$$\delta(\theta_i \mid \theta_j) = (\theta_j/\theta_i) - 1 - \log(\theta_j/\theta_i)\,. \tag{68}$$

Tal com esperavem, $d(\theta_i, \theta_j)$ és una funció simètrica, no negativa còncava que assoleix el seu mínim valor zero si i només si $\theta_i = \theta_j$. L'estimador intrínsec de la taxa de risc és $\theta^*(n, s)$ que minimitza la pèrdua final de referència esperada,

$$\overline{d}(\widetilde{\theta} \mid n, s) = n \int_0^\infty d(\widetilde{\theta}, \theta)\ \mathrm{Ga}(\theta \mid n, s)\, d\theta\,. \tag{69}$$

Amb una aproximació molt bona ($n > 1$), aquesta és $\theta^*(n, s) \approx (2n - 1)/2s$, la mitjana aritmètica entre la mitjana final de referència i la moda final de referència, bastant proper a la mediana final de referència. Amb les dades disponibles, aquesta aproximació dóna $\theta^* \approx 0{,}5893$, mentre que el valor exacte, calculat per minimització numèrica, és $\theta^* = 0{,}5899$. Cal observar que, ja que l'estimació intrínseca és un procediment invariant, l'estimació intrínseca de qualsevol funció $\phi(\theta)$ de la taxa de risc serà simplement $\phi(\theta^*)$.

Test d'hipòtesis. Un criteri d'excel·lència en el sector industrial descrit estableix la producció de primer ordre com aquella que té taxa de risc menor que 0,4, amb la qual cosa el temps de vida esperat és més gran de 2.500 hores. El departament comercial estava interessat a saber si les dades obtingudes eren *compatibles* amb la hipòtesi que la taxa de risc real de la producció de l'empresa era així de petita. La probabilitat final de referència corresponent $\Pr[\theta < 0{,}4 \mid D] = \int_0^{0,4} \mathrm{Ga}(\theta \mid n, s)\, d\theta = 0{,}033$, proporciona una resposta directa, ja que suggereix que la taxa de risc de la producció real podria, possiblement, estar al voltant de 0,4, però que és poc probable que, en realitat, sigui menor.

Pressionats per a donar una mesura quantitativa de la compatibilitat de les dades amb el valor *exacte* $\theta = \theta_0 = 0{,}4$, l'estadístic va donar la discrepància intrínseca esperada $\overline{d}(\theta_0 \mid n, s)$ d'acceptar θ_0 com un substitut del valor vertader de θ sobre la base de les dades (n, s) per a avaluar (69) en $\widetilde{\theta} = \theta_0$. Es va recordar que el valor esperat de $\overline{d}(\theta_0 \mid D)$ sota mostreig repetit és precisament igual a 1 quan $\theta = \theta_0$, i que un valor més gran que $\overline{d}(\theta_0 \mid D)$ indica una forta evidència en contra de θ_0. A més, amb el llenguatge habitual de l'enginyeria, l'estadístic va explicar que els valors de $\overline{d}(\theta_0 \mid D) = d^*$ indicaven,

per a $d^* = 2{,}5, 5{,}0$ o $8{,}5$, un nivell d'evidència contra $\theta = \theta_0$ comparable a l'evidència contra una mitjana 0 que seria proporcionada per a una observació normal x que fou, respectivament, 2, 3 o 4 desviacions estàndards de zero. Com s'indica a la figura 7, els valors de θ_0 més grans que 1,170 o més petits que 0,297 serien convencionalment rebutjats per un criteri normal «3σ». El valor real per a θ_0 fou $\overline{d}(0{,}4 \mid D) = 2{,}01$ (equivalent a $1{,}73\sigma$ sota normalitat). Així, malgrat que hi havia alguna evidència que suggeria que era probable que θ fos més gran que 0,4, no es podria rebutjar el valor exacte $\theta = 0{,}4$ exclusivament en vista a la informació proporcionada per les dades D.

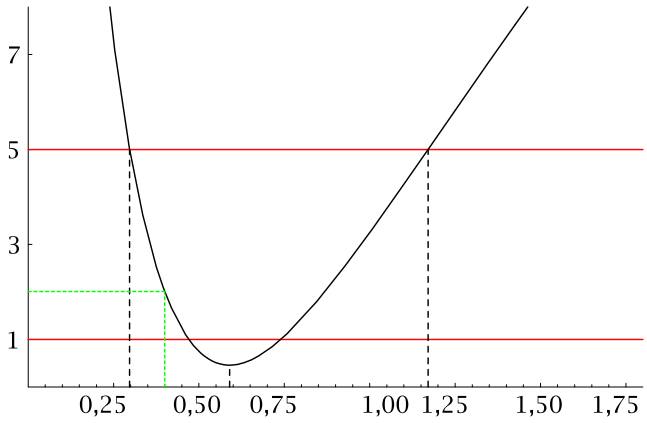

FIGURA 7: Pèrdua intrínseca final de referència esperada quan s'accepta θ_0 com un substitut del valor vertader de θ. El mínim s'assoleix en l'estimador intrínsec $\theta^* = 0{,}590$. Els valors de θ fora de l'interval $(0{,}297, 1{,}170)$ serien convencionalment rebutjats.

6.2 Anàlisi de sensibilitat

Malgrat ser conscients que aquesta informació no podria ser utilitzada en el report preparat per a l'auditoria del client, la direcció de l'empresa estava interessada a formalitzar el coneixement intuïtiu dels enginyers per a adjuntar més informació sobre el temps de vida real dels seus productes. Això es va fer explorant les conseqüències sobre l'anàlisi de *a*) introduir la informació sobre el procés que els enginyers consideraven «més enllà d'un dubte raonable» i *b*) introduir una «millor suposició informada» basada en la seva experiència amb el producte. Els resultats, analitzats més endavant i encapsulats a la figura 8, proporcionen una anàlisi de la sensibilitat de les inferències finals sobre θ a canvis en la informació inicial.

Informació inicial limitada. Preguntats pel consultor estadístic, els enginyers de producció varen afirmar que sabien per experiències passades que el temps de vida mitjà $E[x]$ hauria d'estar al voltant de 2.250 hores, i que possiblement aquesta mitjana no podria ser més gran que 5.000 o més petit que 650. Atès que $\mathrm{E}[x \mid \theta] = \theta^{-1}$, aquestes afirmacions poden ser expressades directament en termes de condicions sobre la distribució inicial de θ; de fet, treballant en milers d'hores, impliquen $\mathrm{E}[\theta] = (2{,}25)^{-1} = 0{,}444$, i $\theta \in \Theta_c = (0{,}20, 1{,}54)$. Per descriure matemàticament aquest coneixement K_1, l'estadístic va utilitzar la corresponent distribució inicial de referència *restringida*, que és la distribució inicial que maximitza la informació desconeguda sobre θ dintre de la classe de distribucions inicials que compleixen aquestes condicions. La distribució inicial de referència restringida a $\theta \in \Theta_c$ i $\mathrm{E}[\theta] = \mu$ és la solució de $\pi(\theta) \propto \theta^{-1} e^{-\lambda\theta}$, subjecta a les restriccions $\theta \in \Theta_c$ i $\int_{\Theta_c} \theta\, \pi(\theta \mid K_1)\, d\theta = 0{,}444$. Amb les dades disponibles, numèricament es va trobar $\pi(\theta \mid K_1) \propto \theta^{-1} e^{-2{,}088\,\theta}$, $\theta \in \Theta_c$. Llavors a partir del teorema de Bayes s'obté que la distribució final corresponent és $\pi(\theta \mid D, K_1) \propto p(D \mid \theta)\, \pi(\theta \mid K_1) \propto \theta^{24} e^{-43{,}69\theta}$, $\theta \in \Theta_c$, una densitat gamma $\mathrm{Ga}(\theta \mid 25, 43{,}69)$ renormalitzada a $\theta \in \Theta_c$, que es representa amb una línia fina a la figura 8. Una comparació amb la distribució final de referència no restringida, descrita per una línia gruixuda, suggereix que, comparada amb la informació proporcionada per les dades, el coneixement addicional K_1 és relativament poc important.

Informació inicial detallada. Quan l'estadístic va continuar preguntant, els enginyers de producció varen conjecturar que el temps de vida mitjà era «segurament» no més gran que 3.000 hores; en demanar-los més precisió, van identificar *segurament* amb un grau 0,95 de creença subjectiva. Treballant amb milers d'hores, això implica que $\Pr[\theta > 3^{-1}] = 0{,}95$. Juntament amb la suposició inicial sobre el temps de vida mitjà que implicava $E[\theta] = 0{,}444$, això va ser suficient per a especificar completament una distribució inicial (subjectiva) $p(\theta \mid K_2)$. Per a obtenir una forma tractable per a aquesta distribució inicial, l'estadístic va utilitzar una rutina numèrica senzilla per a ajustar una distribució gamma restringida a aquelles dues condicions, i va trobar $p(\theta \mid K_2) \propto \mathrm{Ga}(\theta \mid \alpha, \beta)$, amb $\alpha = 38{,}3$ i $\beta = 86{,}3$. A més, l'estadístic va derivar la distribució *inicial predictiva* corresponent $p(x \mid K_2) = \mathrm{Eg}(x \mid \alpha, \beta)$ i va trobar que la distribució inicial obtinguda $p(\theta)$ implicava, per exemple, que $\Pr[x > 1 \mid K_2] = 0{,}64$, $\Pr[x > 3 \mid K_2] = 0{,}27$, i $\Pr[x > 10 \mid K_2] = 0{,}01$, i per tant, les proporcions d'ítems amb un temps de vida per sobre d'1, 3 i 10 milers d'hores eren, respectivament, 64 %, 27 % i 1 %. Els enginyers varen declarar que aquells nombres eren coherents amb llur experiència i d'aquí, l'estadístic va acceptar $p(\theta) = \mathrm{Ga}(\theta \mid 38{,}3, 86{,}3)$, representada amb una línia de punts a la figura 8, com una descripció raonable de les seves creences inicials. Utilitzant el teorema de Bayes, la densitat final que correspon a la inicial $\mathrm{Ga}(\theta \mid \alpha, \beta)$ és $p(\theta \mid D) = p(\theta \mid n, s) \propto \theta^n e^{-\theta s}\, \theta^{\alpha-1} e^{-\beta\theta} \propto \theta^{\alpha+n-1} e^{-(\beta+s)\theta}$, el nucli d'una

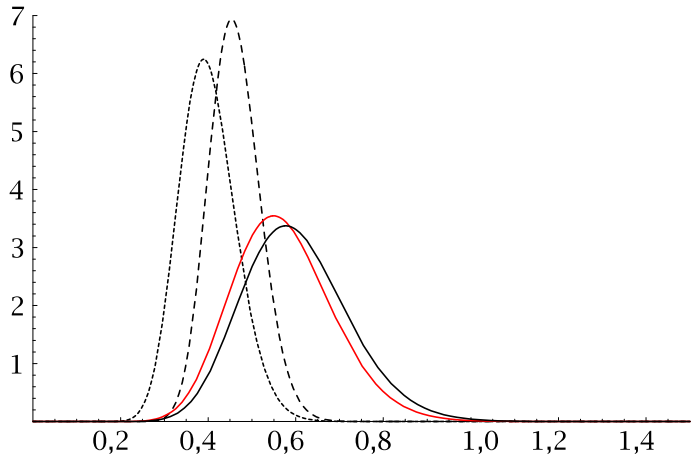

FIGURA 8: Densitats de probabilitat de la taxa de risc θ. Densitat inicial subjectiva (línia de punts), densitat final subjectiva (línia discontínua), final de referència parcialment informativa (línia fina) i final de referència convencional (línia gruixuda).

densitat gamma, de manera que

$$p(\theta \mid D) = \text{Ga}(\theta \mid \alpha + n, \beta + s), \quad \theta > 0\,. \tag{70}$$

Així, la distribució final, que combina el coneixement inicial dels enginyers, K_2 i les dades D va resultar ser $p(\theta \mid D, K_2) = \text{Ga}(\theta \mid 63{,}3, 127{,}8)$, representada amb una línia discontínua a la figura 8. A la figura 8 s'observa fàcilment que les 25 observacions contingudes en les dades analitzades no representen un augment important en la informació sobre la que inicialment tenien els enginyers de producció, malgrat que la distribució final està, de fet, més concentrada que la distribució inicial, i que està desplaçada envers els valors de θ suggerits per les dades. La direcció de l'empresa no va poder usar aquesta informació combinada a l'auditoria però, si confiava en els seus enginyers de producció, va ser aconsellada d'utilitzar $p(\theta \mid D, K_2)$ per a entendre'n millor el procés de producció, o per a dissenyar polítiques per a intentar millorar-ne el rendiment.

7 Discussió i altres aspectes

En escriure un article ampli sempre és difícil decidir què cal deixar fora. En aquest article ens hem concentrat en els conceptes bàsics del paradigma bayesià; els temes metodològics que, amb disgust, s'han omès, inclouen disseny d'experiments, mostreig, models lineals i mètodes seqüències. El lector pot

consultar la bibliografia per a més informació. Aquesta secció final revisa breument els principals arguments per a l'aproximació bayesiana i inclou comentaris sobre altres temes que no s'han discutit amb més detall per limitacions d'espai.

7.1 Coherència

En utilitzar distribucions de probabilitat per a mesurar *totes* les incerteses en un problema, el paradigma bayesià redueix la inferència estadística a probabilitats aplicades, i d'aquesta manera assegura la coherència de les solucions proposades. No hi ha necessitat d'esbrinar, cas per cas, quan la solució a un problema particular és lògicament correcta: un resultat bayesià és només una conseqüència *matemàtica d'hipòtesis completament especificades* i, per tant, llevat que se'n faci un error lògic en la derivació, no pot ser formalment erroni. Al contrari, els mètodes estadístics convencionals estan plens de contreaexemples, com els estimadors negatius de quantitats positives, regions de q-confiança ($q < 1$) que consisteixen en tot l'espai de paràmetres, conjunts buits de solucions «escaients» i respostes incompatibles de metodologies alternatives, simultàniament recolçades per la teoria.

L'aproximació bayesiana demana, però, l'especificació d'una distribució de probabilitat (inicial) sobre l'espai de paràmetre. Sovint s'afirma la frase «una distribució inicial no existeix per a aquest problema» per a justificar l'ús de mètodes no bayesians. Però el teorema general de representació *prova l'existència* d'aquesta distribució sempre que suposem les observacions intercanviables (i si hem assumit una mostra aleatòria, llavors, *a fortiori*, estem assumint que són intercanviables). Ignorar aquest fet matemàtic, i actuar com si una distribució inicial no existís només perquè no és fàcil d'especificar, és matemàticament similar a treballar amb un sistema d'equacions diferencials com si no existís una solució, *un cop s'ha demostrat que una solució existeix*, només perque una solució explícita no és fàcil de trobar.

7.2 Objectivitat

Generalment s'accepta que qualsevol anàlisi estadística és subjectiva, és vol dir que sempre està condicionada a les hipòtesis assumides (sobre l'estructura de les dades, sobre el model de probabilitat, sobre l'espai de resultats) i aquestes hipòtesis, malgrat estar segurament ben fonamentades, són definitivament una elecció *subjectiva*. Llavors, és obligatori formular totes les hipòtesis molt explícitament.

Els usuaris dels mètodes estadístics convencionals poques vegades discuteixen la fonamentació matemàtica de l'aproximació bayesiana, però afirmen ser capaços de produir respostes «objectives» en contrast amb els elements possiblement subjectius involucrats en l'elecció de la distribució inicial.

De fet, els mètodes bayesians demanen l'elecció d'una distribució inicial, i els crítics de l'aproximació bayesiana sistemàticament assenyalen que en moltes situacions importants, incloent-hi informes científics i presa pública de

decisions, els resultats deuen dependre exclusivament de dades documentades que puguin ser objecte d'escrutini independent. Per descomptat, això és veritat, però aquests crítics prefereixen ignorar que aquest cas particular és cobert per l'aproximació bayesiana mitjançant l'ús de distribucions inicials de *referència* que *a*) es dedueixen matemàticament del model de probabilitat acceptat (i, d'aquí, són «objectius» fins al mateix punt que l'elecció del model ho pugui ser) i, *b*) per construcció, produeixen distribucions de probabilitat final que, donat el model de probabilitat acceptat, *només* conté la informació sobre els seus valors que les dades poden proporcionar i, *opcionalment*, qualsevol altra informació contextual sobre la qual hi pugui haver acord universal.

Un altre aspecte relacionat amb l'objectivitat és el del sentit operatiu de les probabilitats finals de referència; l'anàlisi del seu comportament sota mostreig repetit proporciona una forma de calibració suggerent. De fet, $\Pr[\boldsymbol{\theta} \in R \mid D] = \int_R \pi(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$, la probabilitat final de referència que $\boldsymbol{\theta} \in R$, és *alhora* una mesura de la incertesa condicional (donat el model assumit i les dades observades D) sobre l'esdeveniment que el valor desconegut de $\boldsymbol{\theta}$ pertany a $R \subset \Theta$, i la proporció límit de les regions que cobririen $\boldsymbol{\theta}$ sota mostreig repetit amb dades «suficientment similars» a D. Sota condicions febles (per garantir el comportament asimptòtic regular), tots els conjunts grans de dades del mateix model són «suficientment similars» entre si mateixos en aquest sentit, i d'aquí, donades aquelles condicions, les regions finals de referència creïbles són *aproximadament* regions de confiança freqüentista incondicionals.

Les condicions perquè aquesta equivalència *incondicional* aproximada funcioni exclouen, però, casos especials importants com aquells on intervenen observacions «extremes» o «rellevants». En situacions molt especials, quan els models probabilístics poden ser transformats en models de posició i escala, hi ha una equivalència incondicional exacta; en aquests casos els intervals finals de referència creïbles són, per a qualsevol grandària mostral, intervals de confiança freqüentistes incondicionals exactes.

7.3 Aplicabilitat

A diferència de molts mètodes estadístics convencionals, que només poden ser aplicats exactament a unes poques situacions relativament simples i simplificades, els mètodes bayesians són (en teoria) totalment generals. De fet, donat un model de probabilitat i una distribució inicial sobre els seus paràmetres, la derivació de distribucions finals és un exercici matemàtic perfectament definit. En particular, els mètodes bayesians no demanen condicions de regularitat particulars del model probabilístic, no depenen de l'existència d'estadístics suficients de dimensió finita, no es recolzen en teoria asimptòtica, i no demanen la derivació de distribucions mostrals, ni (*a fortiori*) l'existència d'un «pivot» estadístic amb distribució mostral independent dels paràmetres.

Però quan s'utilitzen en models complexos amb molts paràmetres, els mètodes bayesians molt sovint necessiten el càlcul d'integrals definides multi-

dimensionals, la qual cosa, durant molts anys, va posar efectivament límits pràctics a la complexitat dels problemes que podien ser manejats. Això ha canviat moltíssim en els darrers anys amb la disponibilitat general de gran potència de càlcul i amb el desenvolupament paral·lel d'estratègies d'integració numèriques basades en simulació, com *importance sampling* o *Markov chain Monte Carlo* (MCMC). Aquests mètodes proporcionen una estructura dintre la qual molts models complexos poden ser analitzats utilitzat programari genèric. MCMC és *integració numèrica amb la utilització de cadenes de Markov*. La integració de Monte Carlo procedeix traient mostres de la distribució i calculant mitjanes per aproximar les esperances. Els mètodes MCMC trauen la mostra demanada fent córrer durant molta estona cadenes de Markov definides de manera adient; els mètodes específics per a construir aquestes cadenes inclouen *Gibbs sampler* i l'algorisme de Metropolis, originat els anys cinquanta en la bibliografia de física estadística. La producció d'algorismes millorats i el desenvolupament d'eines adients de diagnòstic per a establir-ne la convergència, contínua essent un tema de recerca molt actiu.

La recerca científica sovint demana l'ús de models que són massa complexos per als mètodes estadístics convencionals. Acabarem aquest article amb una breu ullada a alguns.

Estructures jeràrquiques. Considerem una situació on es pren un nombre possiblement variable d'observacions n_i, $\{\boldsymbol{x}_{ij}, j = 1, \ldots, n_i\}$, $i = 1, \ldots, m$, en cadascun de m subconjunts internament homogenis d'una població. Per exemple, una empresa escull m línies de producció per a inspeccionar, i es seleccionen a l'atzar n_i ítems entre els produïts per la línia i, de manera que $\boldsymbol{x}_{ij}$ és el resultat de les mesures fetes a l'ítem j de la línia de producció i. Un altre exemple és el següent: per a estudiar el metabolisme de diverses espècies animals, es capturen animals i se'ls treu una mostra de sang abans de deixar-los anar un altre cop; es repeteix el procediment al mateix habitat durant un temps, de manera que alguns animals són recapturats diverses vegades, i $\boldsymbol{x}_{ij}$ és el resultat de l'anàlisi de la j-èsima mostra de sang de l'animal i. En aquestes situacions, sovint és adient suposar que les n_i observacions de la subpoblació i són intercanviables i, per tant, poden ser tractades com una mostra aleatòria d'un model $p(\boldsymbol{x} \mid \boldsymbol{\theta}_i)$ indexat per un paràmetre $\boldsymbol{\theta}_i$ que depèn de la subpoblació observada, i també podem suposar que els paràmetres que retolen les subpoblacions són intercanviables, de manera que $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$ pot ser tractada com una mostra aleatòria d'una distribució $p(\boldsymbol{\theta} \mid \boldsymbol{\omega})$. Així, el model complet *jeràrquic* que s'assumeix que ha generat les dades observades $D = \{\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{mn_m}\}$ és de la forma

$$p(D \mid \boldsymbol{\omega}) = \int_{\Theta^m} \left[\prod_{j=1}^{n_i} p(\boldsymbol{x}_{ij} \mid \boldsymbol{\theta}_i) \right] \left[\prod_{i=1}^{m} p(\boldsymbol{\theta}_i \mid \boldsymbol{\omega}) \right] \left[\prod_{i=1}^{m} d\boldsymbol{\theta}_i \right] . \qquad (71)$$

D'aquí, sota el paradigma bayesià, una família de models probabilístics convencionals, posem $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, i una distribució inicial «estructural» adient

$p(\boldsymbol{\theta} \mid \boldsymbol{\omega})$, pot ser combinada de manera natural per a reduir un model complex versàtil $\{p(D \mid \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$ l'anàlisi del qual és sovint fora de l'abast de l'estadística convencional. La solució bayesiana només demana l'especificació d'una distribució inicial $p(\boldsymbol{\omega})$, l'ús del teorema de Bayes per a obtenir la corresponent distribució final $p(\boldsymbol{\omega} \mid D) \propto p(D \mid \boldsymbol{\omega})\, p(\boldsymbol{\omega})$, i la realització de les transformacions de probabilitat adients per a derivar les distribucions finals de la quantitat d'interès (que poden també ser funcions de $\boldsymbol{\omega}$, funcions de les $\boldsymbol{\theta}_i$, o funcions d'observacions futures). Com en qualsevol altra anàlisi bayesiana, la distribució inicial $p(\boldsymbol{\omega})$ cal que descrigui el coneixement disponible sobre $\boldsymbol{\omega}$; si no n'hi ha de disponible, o si cal una anàlisi objectiva, es pot usar una funció inicial de referència adient $\pi(\boldsymbol{\omega})$.

Informació contextual. En molts problemes d'inferència estadística, es disposa d'informació contextual objectiva sobre els valors dels paràmetres en la que hi ha un acord universal. Aquesta informació és típicament molt difícil de manejar dintre de l'estadística convencional, però s'incorpora de manera trivial en una anàlisi bayesiana simplement restringint la distribució inicial a la classe $\{\mathcal{P}\}$ de distribucions inicials que són compatibles amb aquesta informació. Per exemple, considerem el problema freqüent en arqueologia d'intentar establir el període d'ocupació $[\alpha, \beta]$ d'un lloc per una cultura passada amb vista a les datacions mitjançant radiocarboni de mostres orgàniques agafades a l'excavació. La datació per radiocarboni no és precisa, de manera que es considera que cada data x_i és una observació d'una distribució normal $N(x \mid \mu(\theta_i), \sigma_i)$, on θ_i és la data real, desconeguda, de la mostra, $\mu(\theta)$ és una corba de calibració internacionalment acceptada, i σ_i és l'error estàndard conegut pel laboratori. Normalment se suposa que les dades reals de calendari $\{\theta_1, \ldots, \theta_m\}$ de les mostres estan uniformement distribuïdes dintre del període d'ocupació $[\alpha, \beta]$; però una evidència estratigràfica indica un ordre parcial, ja que si la mostra i va ser obtinguda a sobre de la mostra j en capes no remogudes, llavors $\theta_i > \theta_j$. Així, si C designa la classe dels valors de $\{\theta_1, \ldots, \theta_m\}$ que compleixen aquestes restriccions, es pot suposar que les dades han estat generades per un model jeràrquic

$$p(x_1, \ldots, x_m \mid \alpha, \beta) = \int_C \left[\prod_{i=1}^{m} N(x_i \mid \mu(\theta_i), \sigma_i^2) \right] (\beta - \alpha)^{-m}\, d\theta_1 \ldots d\theta_m\,. \quad (72)$$

Sovint, la informació contextual també indica una fita inferior absoluta α_0 i una fita superior absoluta β_0 per al període investigat, de manera que $\alpha_0 < \alpha < \beta < \beta_0$. Si no hi ha més informació documentada disponible, cal utilitzar la inicial de referència restringida corresponent per a la quantitat d'interès, $\{\alpha, \beta\}$, que és $\pi(\alpha, \beta) \propto (\beta - \alpha)^{-1}$ quan $\alpha_0 < \alpha < \beta < \beta_0$ i zero en cas contrari. La final de referència corresponent $\pi(\alpha, \beta \mid x_1, \ldots, x_m) \propto p(x_1, \ldots, x_m \mid \alpha, \beta)\, \pi(\alpha, \beta)$ resumeix tota la informació disponible sobre el període d'ocupació.

Informació covariant. En els darrers trenta anys, els models de regressió lineals i no lineal han estat analitzats des d'un punt de vista bayesià amb nivells creixents de sofisticació. Aquests van des de l'elemental anàlisi bayesiana objectiva d'estructures de regressió lineal simple (en paral·lel a les seves contraparts freqüentistes) a la sofisticada anàlisi de sèries temporals involucrades en predicció dinàmica que sovint utilitza estructures jeràrquiques complexes. Aquest camp és excessivament gran per a ser descrit en aquest article, però la bibliografia conté algunes referències rellevants.

Crítica de models. S'ha posat molt èmfasi en el fet que *qualsevol* anàlisi estadística és condicional a les hipòtesis acceptades del model de probabilitat que es pressuposa que han generat les dades. En els darrers anys s'ha fet un gran esforç per a desenvolupar procediments bayesians per a la *crítica de models* i la *selecció de models*. Molts d'aquests procediments són elaboracions sofisticades dels descrits a la secció 4.2 sota la capçalera de «test d'hipòtesis». Un altre cop, aquest és un tema massa llarg per a ser revisat aquí, però algunes referències clau estan incloses a la bibliografia.

Agraïments

L'autor està en deute amb molts col·legues pels seus suggeriments a versions inicials d'aquest article; però mereixen ser especialment agraïts els comentaris detallats que van fer Dennis Lindley, Jennifer Pittman i Reinhard Viertl.

Referències

[1] BERGER, J. O. *Statistical Decision Theory and Bayesian Analysis*, Berlin: Springer, 1985. [Un recull complet de mètodes bayesians que posa èmfasi en els aspectes de la teoria de la decisió].

[2] BERNARDO, J. M. «Expected information as expected utility». *Ann. Statis.*, 7 (1979), 686–690. [Estableix la inferència estadística com un problema de decisió amb una funció d'utilitat basada en la informació].

[3] BERNARDO, J. M. «Noninformative priors do not exist». *J. Statist. Planning and Inference*, 65 (1997), 159–189 (with discussion). [Una anàlisi no tècnica de la polèmica sobre estadística bayesiana objectiva].

[4] BERNARDO, J. M. «Nested hypothesis testing: The Bayesian reference criterion». *Bayesian Statistics* 6 (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds). Oxford: University Press (1999), 101–130 (amb discussió). [Una aproximació als tests d'hipòtesis puntuals des del punt de vista de la teoria de la decisió].

[5] BERNARDO, J. M.; RAMÓN, J. M. «An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters». *The Statistician*, 47 (1998), 1–35. [Una introducció elemental a l'anàlisi bayesiana objectiva].

[6] BERNARDO, J. M.; SMITH, A. F. M. *Bayesian Theory*, Chichester: Wiley, 1994. [Un recull complet per a graduats de conceptes i resultats teòrics d'estadística bayesiana amb una bibliografia extensa].

[7] BERNARDO, J. M.; BERGER, J. O.; DAWID, A. P.; SMITH, A. F. M. (ed.) *Bayesian Statistics 6*. Oxford: University Press, 1999. [Actes del 6th Valencia International Meeting on Bayesian Statistis. Els congressos de València, que se celebren cada quatre anys, proporcionen visions de conjunt definitives de la recerca actual dins del paradigma bayesià].

[8] BERRY, D. A. *Statistics, a Bayesian Perspective*. Belmont, CA: Wasdsworth, 1996. [Una introducció molt bona a l'estadística bayesiana des d'un punt de vista subjectiu].

[9] BOX, G. E. P.; TIAO, G. C. *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley, 1973. [Un recull excel·lent de mètodes bayesians objectius en problemes estadístics estàndard].

[10] DEGROOT, M. H. *Optimal Statistical Decisions*, Nova York: McGraw-Hill, 1970. [Un recull complet sobre la teoria de la decisió bayesiana i la inferència bayesiana amb un tractament rigorós dels fonaments].

[11] EFRON, B. «Why isn't everyone a Bayesian?». *Amer. Statist.*, 40 (1986), 1–11 (amb discusió). [Un bon exemple de la polèmica entre les aproximacions bayesianes i les no bayesianes a l'estadística].

[12] DE FINETTI, B. *Teoria delle Probabilité*, Torí: Einaudi. Traduïda a l'anglès el 1975, *Theory of Probability*. Chichester:Wiley. [Un llibre excepcional sobre probabilitat i estadística des d'un punt de vista subjectiu].

[13] GEISSER, S. *Predictive Inference: an Introduction*. Londres: Chapman and Hall, 1993. [Un recull comparatiu de mètodes de predicció freqüentistes i bayesians objectius].

[14] GELFAND, A. E.; SMITH, A. F. M. «Sampling based approaches to calculating marginal densities». *J. Amer. Statist. Assoc.*, 85 (1990), 398–409. [Un article elemental excel·lent sobre tècniques d'integració numèrica basades en simulació en el context d'estadística bayesiana].

[15] GELMAN, A.; CARLIN, J. B.; STERN, H.; RUBIN, D. B. *Bayesian Data Analysis*. Londres: Chapman and Hall, 1995. [Un tractament general de l'anàlisi de dades bayesiana que posa èmfasi en les eines computacionals].

[16] GILKS, W. R.; RICHARDSON, S.; SPIEGELHALTER, D. J. *Markov Chain Monte Carlo in Practice*. Londres: Chapman and Hall, 1996. [Una introducció excel·lent als mètodes MCMC i llurs aplicacions].

[17] KASS, R. E.; RAFTERY, A. E. «Bayes factors». *J. Amer. Statist. Assoc.* 90 (1995), 773–795. [Una revisió molt interesant dels mètodes de factor de Bayes per a tests d'hipòtesis].

[18] LINDLEY, D. V. *Bayesian Statistics, a Review*. Filadèlfia, PA: SIAM, 1972. [Una revisió general molt intel·ligent de tot el tema fins a la dècada dels anys setanta que posa èmfasi en la consistència interna].

[19] LINDLEY, D. V. «The 1988 Wald memorial lecture: The present position in Bayesian Statistics». *Statist. Sci.*, 5 (1990), 44–89 (amb discussió). [Un recull informatiu del paradigma bayesià i la seva relació amb altres actituds envers la inferència].

[20] LINDLEY, D. V. «The philosophy of statistics». *The Statistician*, 49 (2000), 293–337 (amb discussió). [Una descripció recent del paradigma bayesià des d'un punt de vista subjectiu].

[21] O'HAGAN, A. *Bayesian Inference*. Londres: Edward Arnold, 1994. [Un bon recull d'inferència bayesiana integrada a la Biblioteca d'Estadística de Kendall].

[22] PRESS, S. J. *Applied Multivariate Analysis: using Bayesian and Frequentist Methods of Inference*. Melbourne, FL: Krieger, 1972. [Un recull comparatiu general de mètodes d'inferència en problemes multivariants freqüentistes i bayesians objectius].

[23] PRESS, S. J.; TANUR, J. M. *The Subjectivity of Scientists and the Bayesian Approach*. Nova York: Wiley, 2001. [Una descripció molt interessant de cóm al llarg de la història les creences preconcebudes de científics famosos varen influir en llurs conclusions científiques].

[24] WEST, M.; HARRISON, P. J. *Bayesian Forecasting and Dynamic Models*. Berlin: Springer, 1989. [Un recull complet excel·lent sobre l'anàlisi bayesiana de sèries temporals].

[25] ZELLNER, A. *An Introduction to Bayesian Inference in Econometrics*. Nova York: Wiley. Reinpressió el 1987; Melbourne, FL: Krieger, 1971. [Una anàlisi bayesiana objectiva detallada de models lineals].

DEPARTAMENT D'ESTADÍSTICA I INVESTIGACIÓ OPERATIVA
FACULTAT DE MATEMÀTIQUES
UNIVERSITAT DE VALÈNCIA
46100 BURJASSOT, VALÈNCIA
jose.m.bernardo@uv.es