HALANDÓSÁGI TÁBLÁK BECSLÉSE BAYESI MÓDSZEREKKEL[1]

GÁL PÉTER

Dolgozatomban az életbiztosítások matematikájának egy alapvető feladatára, a halandósági táblák szerkesztésére mutatok be lehetséges módszereket. Megközelítésem kiindulópontja a bayesi statisztika. A statisztikai következtetéselmélet e viszonylag fiatal vonulata jelentőségéhez és sokrétű alkalmazhatóságához képest kevéssé ismert, ezért a dolgozat első fejezetében röviden bemutatom elvi alapjait és a későbbiek során használt néhány módszerét és tulajdonságát. A második fejezet a címben szereplő fontos gyakorlati problémára tett megoldási javaslatokkal hívja fel a figyelmet a megközelítés hatékonyságára, flexibilitására. Szeretnék köszönetet mondani a dolgozat elkészítésében nyújtott segítségéért Hunyadi Lászlónak, a BKE egyetemi tanárának és Szabó Józsefnek, a Winterthur Biztosító Rt. vezető matematikusának.

1 A bayesi statisztika alapjai

"Két esemény együttes bekövetkezésének valószínűsége egyenlő az első bekövetkezésének valószínűsége szorozva annak valószínűségével, hogy a második bekövetkezik, feltéve, hogy az első bekövetkezik."[2] Ezt az állítást közel két és fél évszázaddal ezelőtt fogalmazta meg az angol matematikus, filozófus Thomas Bayes tiszteletes, s erre az állításra épült később a klasszikus valószínűségszámítás Bayes-tétele.

Thomas Bayes tiszteletes
(1702-1761)

[1]Beérkezett: 1998. március 28.

[2]Ld. [1].

A tétel a huszadik század közepén egy heves vitákat kiváltó statisztikai megközelítés alapjául szolgált. A bayesi statisztika szemléletében és eszközeiben sok helyütt szakított a klasszikus elmélettel. A bayesi elemzések során következtetéseinket a klasszikus úttal ellentétben nem csak a mintából nyerhető információk alapján hozzuk. Lehetőség nyílik szubjektív vélemények egzakt kezelésére, szubjektív valószínűségek használatára. Nem feltételezzük az ismételt mintavétel lehetőségét, a rendelkezésre álló egyetlen minta alapján dolgozunk. Továbbá —talán a legjelentősebb szemléletbeli különbség, hogy— a bayesi statisztika becslésének tárgya nem rögzített érték, hanem véletlen változó és a becslés e változó valószínűség-eloszlásának meghatározását jelenti.

A bayesi megközelítést a külső információk felhasználása és a szubjektív valószínűségek használata miatt sok támadás érte (éri). Hívei éppen abban látják erejét, hogy véleményük szerint egyrészt a statisztikus elemzése során mindig felhasznál külső információt (pl. sikertelen kísérletek elvetése) és az elemzés átláthatóságát növeli, ha ezek formálisan is megjelennek a modellben; másrészt a bayesi megközelítésben elképzelhető a külső információk hiánya is, azaz a klasszikus mintavételi statisztika speciális esetként jelenik meg benne. Továbbá a bayesi elméletben is találkozhatunk a klasszikus statisztika alapfeladatainak, a pont- és intervallumbecslés, a hipotézisvizsgálat analogonjaival.

A bayesi módszerek igencsak számításigényes eljárásokat használnak, ezért nem meglepő, hogy e megközelítés népszerűsége a számítástechnika robbanásszerű fejlődését követte. Manapság már sok területen alkalmazható a hagyományos módszerek mellett vagy helyett.

A Bayes-tétel, a prior és a posterior fogalma

A bayesi megközelítés alapja a Bayes-tétel —melyet gyakran a fordított valószínűség tételének is neveznek— folytonos valószínűségi változókra a következőképpen formalizálható:

$$f(\mathbf{y},\theta) = f(\mathbf{y}\,|\,\theta)f(\theta) = f(\theta\,|\,\mathbf{y})f(\mathbf{y})\,, \tag{1}$$

azaz

$$f(\theta|\mathbf{y}) = \frac{f(\theta)f(\mathbf{y}\,|\,\theta)}{f(\mathbf{y})}\,, \tag{2}$$

ahol $\mathbf{y}$ megfigyelések véletlen vektora, θ szintén valószínűségi változónak tekintett paramétervektor, $f(\mathbf{y},\theta)$, $f(\mathbf{y}\,|\,\theta)$ és $f(\theta\,|\,\mathbf{y})$ pedig az együttes illetve feltételes sűrűségfüggvényeket jelölik. Mindezt a következő formában is írhatjuk:

$$f(\theta\,|\,\mathbf{y}) \propto f(\theta)f(\mathbf{y}\,|\,\theta)\,, \tag{3}$$

ahol $\propto$ arányosságot jelöl. Az arányossági tényező nyilván $1/f(\mathbf{y})$, mely egy konkrét mintát tekintve konstans. Az $f(\theta)$-t a θ paramétervektor a priori sűrűségfüggvényének vagy röviden priorjának nevezzük. Az $f(\theta)$ az a posteriori sűrűségfüggvény vagy posterior, $f(\mathbf{y}\,|\,\theta)$ pedig a mintavételi statisztikából ismert likelihood függvény. (3) tehát szavakkal a következőképpen fogalmazható meg: a posterior sűrűségfüggvény arányos a prior és a likelihood függvény szorzatával.

A posterior tartalmaz mintán kívüli, előzetes információt, mely a prior révén jelenik meg benne és mintából származó információt, melyet a likelihood függvény közvetít. A bayesi elemzések során a posterior sűrűségfüggvényt használjuk a paraméterekre vonatkozó következtetések levonásához.

A prior

A prior sűrűségfüggvény, $f(\theta)$ modellünk paramétereire vonatkozó előzetes ismereteinket reprezentálja. A bayesi elemzés egyik döntő és egyáltalán nem magától értetődő mozzanata a megfelelő, a külső információkat vagy azok hiányát a lehető leghűebben közvetítő prior megválasztása. Fontos, hogy a prior matematikailag egyszerűen kezelhető legyen. További követelmény, hogy független legyen a mintától.

Az a priori információ származhat korábbi mintákkal, adatokkal végzett elemzésekből, ilyenkor adatokon alapuló információról beszélünk. Máskor elméleti megfontolásokról, szubjektív feltételezésekről, szakértői megérzésekről van szó: ezek a nem adatokon alapuló priorok. E két fajta információ természetesen nem mindig választható szét egyértelműen. Gyakran előfordul az az eset, amikor csak kevés vagy semmilyen előzetes információ nem áll rendelkezésünkre. Az ilyen prior megalkotása sem nyilvánvaló, sőt a bayesi következtetéselmélet egyik legnehezebb és legtöbb vitát kiváltó kérdése.

Szólni kell még a priorok egy fontos családjáról, a természetes konjugált prioroknak nevezett sűrűségfüggvényekről. Az ilyen priorok használatát a lánctulajdonságnak való megfelelés igénye motiválja. Szerencsés tulajdonsága ugyanis a priornak, ha valamely gyakori likelihood függvénnyel való szorzataként előálló posterior egy következő következtetési lépésben megfelelő priorként szerepelhet. A természetes konjugált priorokból származtatott posterior típusa megegyezik a prioréval, így nyilván rendelkezik a lánctulajdonsággal. Példa természetes konjugált priorra a binomiális eloszlás p paraméterére feltételezett béta eloszlás. Ismert, hogy ekkor a likelihood függvény és a prior sűrűségfüggvény is $P^k(1-P)^{n-k}$-val arányos és ilyen függvények szorzata nyilván hasonló alakú. További természetes konjugált családok a normális-normális vagy a Poisson-gamma modellek.

A likelihood függvény

A bayesi elemzésben a likelihood függvény hordozza a mintából származó információkat. Gyakran feltételezzük, hogy a mintaelemek független azonos feltételes eloszlásból származnak. Ilyenkor a likelihood függvény felírható az alábbi módon:

$$f(\mathbf{y}\,|\,\theta) \propto \prod_i f(y_i\,|\,\theta)\,. \tag{4}$$

Érdemes felhívni a figyelmet a likelihood függvény egy fontos tulajdonságára, miszerint a mintából nyerhető összes információt tartalmazza, azaz "mindent elmond, amit a minta mondhat".[3]

Következtetés több mintából

Tegyük fel, hogy a θ paramétervektor vizsgálata során rendelkezésünkre áll az $f(\theta)$ prior mellett két független kísérletből származó minta: $\mathbf{y}_1$, $\mathbf{y}_2$. Ilyen esetben a posterior elkészítésekor a következőképpen járhatunk el: képezzük a (3) szabály szerint az első minta alapján az

$$f(\theta \,|\, \mathbf{y}_1) \propto f(\theta) f(\mathbf{y}_1 \,|\, \theta) \tag{5}$$

posteriort, majd ezt tekintve prior sűrűségfüggvénynek a második mintát felhasználva számíthatjuk a két mintára támaszkodó posteriort:

$$f(\theta \,|\, \mathbf{y}_1, \mathbf{y}_2) \propto f(\theta \,|\, \mathbf{y}_1) f(\mathbf{y}_2 \,|\, \theta) \,. \tag{6}$$

Az előző két arányosság összevetésével a posterior egy fontos tulajdonságára figyelhetünk fel:

$$f(\theta \,|\, \mathbf{y}_1, \mathbf{y}_2) \propto f(\theta) f(\mathbf{y}_1 \,|\, \theta) f(\mathbf{y}_2 \,|\, \theta) \,. \tag{7}$$

Az $f(\mathbf{y}_1 \,|\, \theta) f(\mathbf{y}_2 \,|\, \theta)$ szorzat nem más, mint a θ paraméter likelihood függvénye az $\mathbf{y}_1$, $\mathbf{y}_2$ minták alapján, azaz

$$f(\theta \,|\, \mathbf{y}_1, \mathbf{y}_2) \propto f(\theta) f(\mathbf{y}_1, \mathbf{y}_2 \,|\, \theta) \,. \tag{8}$$

Azt találtuk tehát, hogy a prior különböző minták alapján való ismételt módosítása azonos eredményre vezet az egyesített mintából számított likelihood függvény segítségével képzett posteriorral. Megmutatható, hogy ez az állítás kettőnél több mintára is érvényes.

A fentiek alapján az újonnan szerzett információk folyamatosan beépíthetők a posteriorba, így az a mindenkor rendelkezésünkre álló összes előzetes és mintából származó információt tartalmazhatja. Nem szükséges tehát az a priori és a korábbi mintainformációk megőrzése és az azokból való újraszámolás.

Paraméterek pontbecslése

A bayesi megközelítés a paraméterek teljes posterior eloszlását szolgáltatja. Ezeket az eloszlásokat leírhatjuk a szokásos statisztikai jellemzőkkel: középértékekkel, szórással, ferdeséggel. Kérdés, hogy a paraméter pontbecslésére az eloszlás mely jellemzőjét érdemes használni. Kézenfekvő megoldás valamely középértéket (várható értéket, móduszt stb.) választani.

Egy másik lehetséges megközelítés valamilyen, a becslésből eredő „veszteség" minimalizálása. Ehhez egy megfelelő veszteségfüggvényt ($L(\theta, \hat{\theta})$) választunk, melynek változói a véletlen változónak tekintett θ és ennek egy

[3] L. J. Savage: Subjective Probability and Statistical Practice, in: The foundations of Statistical Inference, Methuen & Wiley, London, New York, 1961 idézi Zellner [6].

adott mintához tartozó becslése ($\hat{\theta} = \hat{\theta}(y)$). A veszteségfüggvény a becslés „rosszságát", a θ értékétől való távolságát hivatott mérni. Ne feledjük, hogy a bayesi megközelítés szerint nincs egyetlen "valódi" θ érték, hanem a paraméterek posterior sűrűségfüggvénye áll rendelkezésünkre. Ezért θ becslésének azt a pontot választjuk, amelyben a veszteségfüggvény posterior várható értéke minimális:

$$\hat{\theta} = \underset{\theta'}{\operatorname{argmin}}\, E\left(L(\theta, \theta')\right) = \underset{\theta'}{\operatorname{argmin}} \int_{R(\theta)} L(\theta, \theta') f(\theta \,|\, \mathbf{y})\, d\theta\,, \tag{9}$$

feltéve, hogy a várható érték véges és a minimum létezik.

A veszteségfüggvényt sokféleképpen megválaszthatjuk, egy gyakori megoldás a kvadratikus,

$$L = (\theta - \hat{\theta})' C (\theta - \hat{\theta})$$

alak, ahol C egy konstans szimmetrikus pozitív definit mátrix. Ilyenkor a veszteségfüggvény várható értéke a következő formára alakítható:[4]

$$\begin{gathered} E\big(L(\theta, \hat{\theta})\big) = E\big((\theta - \hat{\theta})' C (\theta - \hat{\theta})\big) = \\ E\Big(\big(\theta - E(\theta) - (\hat{\theta} - E(\hat{\theta}))\big)' C \big(\theta - E(\theta) - (\hat{\theta} - E(\hat{\theta}))\big)\Big) = \\ E\Big((\theta - E(\theta))' C (\theta - E(\theta)) + (\hat{\theta} - E(\theta))' C (\hat{\theta} - E(\theta))\Big)\,. \end{gathered} \tag{10}$$

Az utolsó várható érték első tagja független $\hat{\theta}$-tól, a második tag, $(\hat{\theta} - E(\theta))' C (\hat{\theta} - E(\theta))$ pedig nem sztochasztikus, és minimumát akkor veszi fel, ha $\hat{\theta} = E(\theta)$. Azt a —nem meglepő— eredményt kaptuk tehát, hogy a posterior várható értéke optimális pontbecslést ad a paraméter értékére (kvadratikus veszteségfüggvény esetén).

Nagymintás tulajdonságok

Ebben a bekezdésben a bayesi posterior sűrűségfüggvények néhány nagymintás tulajdonságát tárgyaljuk. A posteriorról tudjuk, hogy a prior és a likelihood függvény szorzatával arányos:

$$f(\theta \,|\, \mathbf{y}) \propto f(\theta) f(\mathbf{y} \,|\, \theta) = f(\theta) e^{\ln f(\mathbf{y} \,|\, \theta)}\,. \tag{11}$$

Tegyük most még fel, hogy $\mathbf{y}$ egy n elemű vektor, azaz n független megfigyelésünk van. Ekkor $\ln f(\mathbf{y} \,|\, \theta)$ általában n-nel arányos,[5] míg $f(\theta)$ független a minta méretétől (mivel független a mintától). Így nagy minták esetén a likelihood függvény dominálja a posterior sűrűségfüggvényt. Zellner[6] igazolja, hogy (általánosan fennálló feltételek mellett) a posterior aszimptotikusan normális eloszlású lesz, a maximum likelihood becsléssel egyenlő várható

[4] Az első lépésben a posterior várható értékét, $E(\theta)$-t levonjuk és hozzáadjuk a veszteségfüggvényhez. Ez a várható értéket nem befolyásolja. A második lépésben kihasználjuk, hogy az $E\Big(\big(\theta - E(\theta)\big)' C \big(\hat{\theta} - E(\theta)\big)\Big)$ keresztszorzatok kiesnek, mivel $E\big(\theta - E(\theta)\big) = 0$.

[5] Lindley [5]

[6] [6] 32. old.

értékkel. Az is megmutatható, hogy a minta növekedésével a posterior szórása csökken.[7] Nagy minták esetén tehát a bayesi posterior sűrűségfüggvény a maximum likelihood becslés szűk környezetében „összpontosul".

Összefoglalás

Az első fejezetben bemutattuk, hogy a bayesi elmélet milyen, a klasszikus felfogástól eltérő elvekkel közelít a statisztika feladataihoz. A bayesi módszerek hatékonyan alkalmazhatók a klasszikus statisztika legtöbb feladatának alternatív megközelítéssel való megoldására, így a gazdasági élet döntéseinek statisztikai megalapozására vagy ezen belül a kifejezetten aktuáriusi problémák megoldására is. A vagyonbiztosítási kalkulációk során előszeretettel alkalmazott megbízhatósági elmélet például előszeretettel alkalmaz bayesi módszereket a megbízhatósági tényező meghatározására.[8] Az alábbiakban az életbiztosítások területén mutatunk be alkalmazásokat.

2 A halandósági táblák problémája

Az életbiztosítások díjait a biztosítótársaságok halandósági táblák alapján határozzák meg. A biztosítók számára fontos, hogy a halandósági táblák pontos adatokat tartalmazzanak, hiszen haláleseti biztosítások esetén a halandóság alulbecslése, elérési típusú biztosítások esetén a halandóság felülbecslése veszteséget okozhat. A biztosító nem törekedhet „biztonsági játékra" a halandóságok szándékos felül-, illetve alulbecslésével, hiszen ez indokolatlanul magas biztosítási díjakat eredményezne, ami a versenyképesség rovására mehet.

A statisztikai hivatalok készítenek az egész népességre vonatkozó halandósági táblákat, ezek azonban a biztosítótársaságok számára több okból sem szolgáltatnak megfelelő adatokat. Egyrészt az életbiztosítást vásárlók eloszlása a társadalomban nyilván nem egyenletes: pl. a legszegényebb rétegek, melyek halandósága az átlagnál magasabb, az átlagnál kisebb arányban kötnek életbiztosítást, így a biztosított állomány halandósági mutatói általában alacsonyabbak, mint a teljes társadalomra jellemző értékek. Ezt figyelembe véve a biztosítótársaságok szokás szerint a teljes népességére meghatározott halandósági rátákat (q_x) egy minden korosztályra megegyező 1-nél kisebb konstanssal szorozzák meg és az így előálló halandósági táblával kalkulálnak.

Ugyancsak megkérdőjelezhetjük az országos halandósági táblák életbiztosítások díjkalkulációjánál való alkalmazásának jogosságát a biztosító ügyfelei között megjelenő antiszelekció miatt: a tapasztalatok szerint kockázati életbiztosítást az összes biztosítást vásárló átlagánál nagyobb, elérési biztosítást pedig az átlagnál kisebb halandóságú ügyfelek kötnek.

[7] [6] 33. old.

[8] Ld. pl. [4].

A biztosítónak tehát a potenciális kockázati, illetve elérési életbiztosítás vásárlók halandósági táblájára van szüksége. Ez a feladat sugallja a bayesi megközelítés alkalmazását. Rendelkezésünkre állnak ugyanis a priori feltételezések a táblára vonatkozóan és egy folyamatosan növekvő minta, a biztosító szerződéseinek állománya. A külső feltételezések között a fenti szociológiai megfontolások és az országos halandósági tábla mint kiindulópont mellett további elvárásokat is megfogalmazhatunk a priorban, például külföldi tapasztalatok alapján. Igényként merülhet fel az is, hogy a halandósági táblában szereplő értékek analitikus függvénnyel kifejezhetők legyenek vagy más kedvező tulajdonsággal rendelkezzenek, például ne "ugráljanak", azaz egymáshoz közeli korosztályok halandósági rátái csak kevéssel térjenek el egymástól. Mindezek az elvárások a megfelelően megválasztott prioron keresztül jelenhetnek meg a modellben.

A bayesi megközelítés alkalmazásának elfogadása még nem jelent egy konkrét receptet a probléma megoldására, csak annyit, hogy feltételezéseinket, külső információinkat a priorba, a mintából származó információt a likelihood függvénybe foglalva vonjuk modellünkbe, és az e két tényező által meghatározott posterior lesz következtetésünk, becslésünk alapja. A továbbiakban láthatjuk, hogy e gondolatmenet alkalmazása többféle különböző módszert és eredményt szolgáltathat. Minden példában a Mellékletben található adatokkal számolunk, melyek egy fiktív biztosítótársaság állományának nagyságát[9] és a bekövetkezett halálesetek számát mutatja kor szerinti bontásban három különböző évben, továbbá a „hagyományos", országos halandósági tábla fent leírt egyszerű módosításával meghatározott halandósági rátákat.

1. módszer: Béta-binomiális modell a halandósági ráták becslésére

Első megoldásunkban a különböző korokban jellemző halandósági rátákat egyenként (korévenként külön-külön), egymástól függetlenül becsüljük meg. Az x éves korú biztosítottak halandósági mutatójának meghatározásához rendelkezésünkre álló információk a következők: az országos halandósági táblában szereplő érték (q_x, erre alapozzuk a priort), a biztosító x éves korú állományának nagysága (l_x) és az ezen állománybeli halálesetek száma (d_x). A halálozások relatív gyakoriságát jelöljük u_x-szel ($u_x = d_x/l_x$), a halandósági ráták keresett becslését pedig w_x-szel. Mivel csak egy évjárattal foglalkozunk, az x indexeket ott, ahol ez nem vezet félreértéshez, a továbbiakban elhagyjuk. A halandósági mutatók kiigazítását rendszeresen (pl. évente) érdemes elvégezni, ezért fontos, hogy a prior lánctulajdonságú legyen. Ha természetes konjugált priort akarunk használni, annak kiválasztásához ismernünk kell a likelihood függvény alakját. Ezért első lépésként most határozzuk meg ezt.

Feltételezhetjük, hogy a biztosító állományában bekövetkező halálozások

[9] A fiktív adatokat úgy állítottam elő, hogy egy létező biztosítótársaság különböző években tapasztalt adataiból képeztem összegeket. Erre azért volt szükség, mert a valós adatok nem publikusak, ugyanakkor így életszerű adatokkal számolhattam.

száma binomiális eloszlást követ, melynek paraméterei a keresett halálozási arány (w) és az állomány nagysága (l). A likelihood függvény tehát:

$$f(d\,|\,w) = \binom{l}{d} w^d (1-w)^{l-d} \,. \tag{12}$$

A következő lépés a prior meghatározása. Szerencsés béta eloszlású priort választani, hiszen tudjuk, hogy az a fenti likelihood függvénnyel természetes konjugált. A béta eloszlás paramétereit úgy választjuk meg, hogy a várható érték q-nak adódjon, azaz a priori információként az országos halandósági rátát használjuk fel. Mivel két paramétert adhatunk meg, meghatározhatjuk még a szórást is. Ezt a különböző korosztályok esetében különböző nagyságúnak tanácsos választanunk, hiszen q_{18} 0,001 nagyságrendű, míg q_{80} 0,1-hez közeli értékű. Használjunk ezért a különböző korokra egy fix r relatív szórású eloszlást. A béta eloszlás α és β paraméterének meghatározásához a következő kétismeretlenes egyenletrendszert kell megoldanunk:

$$\begin{aligned} \frac{\alpha}{\alpha+\beta} &= q \\ \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} &= (rq)^2 \,. \end{aligned} \tag{13}$$

Innen

$$\begin{aligned} \alpha &= \frac{1-q-r^2q}{r^2} \\ \beta &= \frac{(1-q-r^2q)(1-q)}{r^2q} \,. \end{aligned} \tag{14}$$

Mindezek alapján a béta eloszlású prior sűrűségfüggvénye már meghatározható:

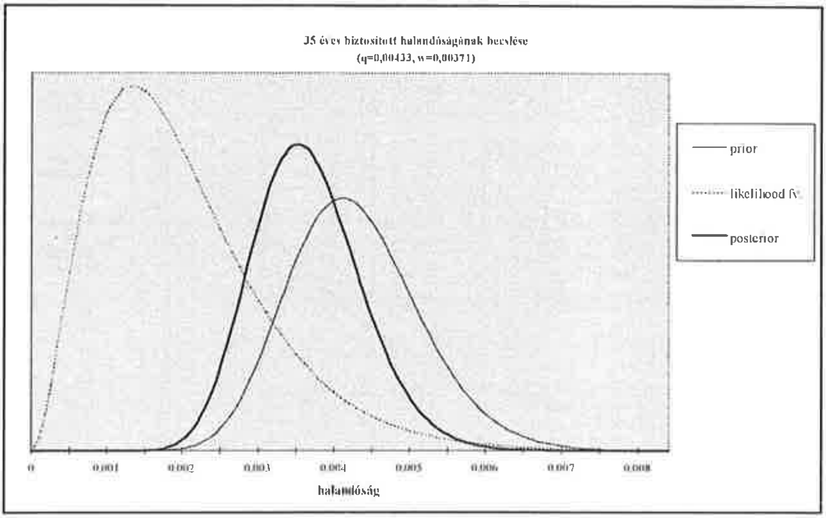
$$f(w) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} w^{\alpha-1}(1-w)^{\beta-1} \,. \tag{15}$$

A posterior pedig (3) alapján a következő alakú:

$$f(w\,|\,d) \propto w^{d+\alpha-1}(1-w)^{l-d+\beta-1} \,, \tag{16}$$

azaz $d+\alpha$ és $\beta+l-d$ paraméterű béta eloszlású. A posterior sűrűségfüggvény alapján meghatározhatjuk az $E(w\,|\,d) = \frac{d+\alpha}{l+\alpha+\beta}$ posterior várható értéket, amely w bayesi pontbecslésének tekinthető.

Nézzük meg konkrét adatainkra, milyen eredményt kapunk a fenti módszerrel. Tekintsük egy 35 éves biztosított halandóságának becslését. A prior meghatározásához felhasználjuk a rendelkezésre álló halandósági tábla megfelelő értékét: $q_{35} = 0.00433$. 20%-os relatív szórást feltételezve a w eloszlását leíró béta eloszlású prior paramétereire (14) alapján $\alpha = 24.89$ és $\beta = 5722.8$ adódik. A likelihood függvény meghatározásához szükségünk van az állomány nagyságára ($l = 1505$ fő) és a halálozások számára ($d = 2$). A fentiek alapján a posterior béta eloszlású 26.89 és 7225.8 paraméterekkel. Várható értéke pedig, melyet w becslésének tekintünk $\frac{26.89}{26.89+7225.8} = 0.00371$. Az *1. ábrán* a prior, a likelihood függvény és a posterior alakja látható.



1. ábra

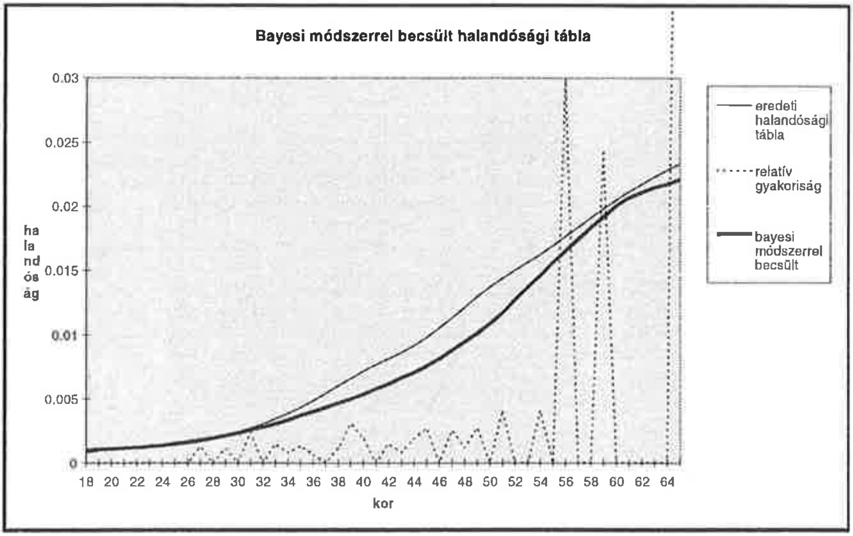
Látható, hogy becslésünk a halandósági táblabeli érték és az ennél jóval kisebb tapasztalt relatív gyakoriság (2/1505 = 0.00133) közé esik. Azt is megfigyelhetjük, hogy a viszonylag kis minta kisebb súllyal szerepel a módosított halandósági érték meghatározásánál.

A választott béta-binomiális modell lánctulajdonságát kihasználva újabb mintákból nyert információinkat egyszerűen beépíthetjük a becslésbe, a most kapott posteriort tekintve priornak. Becsüljük meg a 35 éves biztosított halandóságát az első becslést és a második évi adatokat felhasználva! Most $\alpha = 24.9$, $\beta = 6692.7$, $l = 3102$, $d = 6$, így $w = 30.9/9819.6 = 0.00315$. Megfigyelhető, hogy tovább közelítettünk a relatív gyakorisághoz, ami indokolt, hiszen most nagyobb, és ezért megbízhatóbb minta alapján becsültünk.

Természetesen a fenti becslés valós helyzetekben való alkalmazásához a körülmények további elemzése is szükséges. Egy fiatal biztosítótársaság[10] esetén például az állomány kicsi, így az empirikus halandósági arányok szórása (és ezért a szomszédos értékek közötti ingadozása) nagyon nagy lehet, és ez nagy ingadozást eredményez a bayesi módon becsült értékekben is. Mivel feltételezhető, hogy a szomszédos korcsoportok halandósága nem nagyon tér el egymástól, ezt a káros jelenséget az eredményül kapott tábla adataira illesztett "sima" görbe használatával részben orvosolhatjuk. Talán a legegyszerűbb simítási módszer mozgóátlag használata. A simítást alkalmazhatjuk a bemenő adatokra (a tapasztalt relatív gyakoriságokat simítjuk) vagy a kimenetre (a becsült halandósági rátákra). A *2. ábra* a fenti módszerrel becsült halandósági mutatók 9 tagú mozgóátlagolással simított értékeit valamint az eredeti halandósági tábla adatait és a relatív gyakoriságokat mutatja:

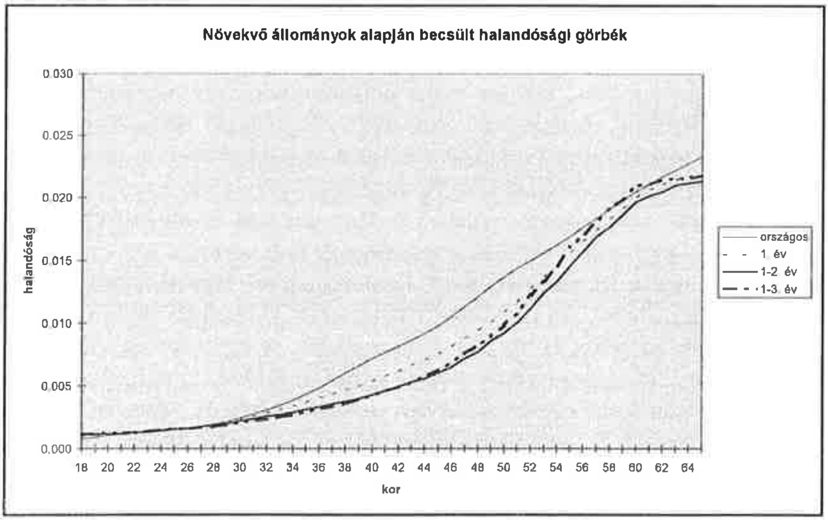
---

[10] Magyarországon a feldogozott adatok mennyisége szempontjából a biztosítótársaságok többnyire „fiatalnak" tekinthetők.



2. ábra

Egy másik probléma, hogy mivel a biztosítás megkötése előtt az ügyfelek egészségi állapotát megvizsgálják, az azonos korú régi és új ügyfelek halandósága is eltérhet egymástól. E csoportok különválasztása tovább csökkenti a minta méretét. Mint láttuk, ez azt eredményezi, hogy a likelihood függvény súlya a posteriorban kicsi lesz, s így becslésünk az a priori feltételezett halandósági táblához közeli értéket szolgáltat majd. Magyarán eljárásunk ilyen esetekben sok új információt nem nyújt.



3. ábra

A minta méretét és így súlyát növelhetjük, ha a következő évek adatait rendre bevisszük a modellbe. A *3. ábra* a kiinduló halandósági tábla mellett az egy, kettő, illetve három év megfigyeléseire támaszkodó bayesi módon becsült, majd kisimított halandósági táblákat tartalmazza. A második illetve harmadik év adatainak bevonásakor a priori információnak már a megelőző évben meghatározott kisimított halandósági táblát tekintettük.

Látható, hogy a becsült görbe az évek múlásával jobban eltérhet az eredeti, országos halandósági táblától, mivel az egyre nagyobb állományok miatt a minta súlya a becslésekben egyre nő.

2. módszer: A halandósági tábla „sima" becslése

A következő, 2. módszer bár a fentivel azonos feladat megoldására szolgál és szintén bayesi elvek alapján közelíti meg a problémát, egyéb, az 1. módszerrel közös vonással nem rendelkezik. Az egyes korokra jellemző halandósági rátákat nem külön-külön, hanem szimultán módon becsüljük. Az eredményként kapott halandósági görbe simaságát nem utólagos „kezeléssel" biztosítjuk, hanem ezt az elvárásunkat már a priorban megfogalmazzuk. Az előző eljárással ellentétben az életbiztosítási halandósági tábla becslése során egyáltalán nem támaszkodunk az országos halandósági táblára. Ezzel egy biztosan torzító elemet távolítunk el a modellből.

Mint azt már a korábbiakban tárgyaltuk, a becsült halandósági táblával szembeni legfontosabb elvárásunk, hogy „sima" legyen, valamint, hogy a mintából meghatározott relatív gyakoriságokkal a lehető leginkább összhangban álljon. Ezt a két elvárt tulajdonságot rendre a priorban és a mintából nyerhető információt közvetítő likelihood függvényben fogalmazzuk meg.

További újdonsága a megközelítésnek, hogy az egyes korosztályokra ($x = 18..65$) jellemző halandósági ráták, mint valószínűségi változók együttes eloszlására teszünk a priori feltevést, illetve szerkesztjük meg a likelihood függvényt.

A priorban szándékozunk megfogalmazni a simaságra vonatkozó elvárásainkat. Ez azt jelenti, hogy a halandósági ráták w_x becsléseinek együttes eloszlását olyannak tételezzük fel, hogy kis valószínűséget tulajdonítunk a szomszédos értékek nagy eltéréseinek. Az eltérések mérésére használhatjuk például az euklideszi távolságot. A prior sűrűségfüggvény megadására az exponenciális függvényt választjuk, mert könnyen kezelhető, monoton függvényre van szükségünk. Látni fogjuk, hogy ez a választás a későbbiek folyamán nagyban leegyszerűsíti majd a számításokat. Legyen tehát a prior:

$$f(w_{18}, \ldots, w_{65}) \propto \exp\left(-\sum_{x=18}^{64}(w_x - w_{x+1})^2\right) . \qquad (17)$$

A likelihood függvényt hasonló elvek alapján szerkesztjük meg. Itt azt kell elérnünk, hogy a w_x becslések az u_x relatív gyakoriságoktól kevéssé térjenek el. Azokat a korosztályokat, melyekben az állomány nagyobb, nyilván nagyobb súllyal kell figyelembe vennünk, mint a jelentősebb mintavételi hibalehetőséggel terhes kevés főből állókat. A likelihood függvényt ezért a következőképpen határozzuk meg:

$$f(u_{18},\dots,u_{65}\,|\,w_{18},\dots,w_{65}) \propto \exp\left(-\sum_{x=18}^{65} ml_x(u_x-w_x)^2\right), \tag{18}$$

ahol az m a minta és a prior súlyozására szolgáló faktor.

A fenti egyszerűen megfogalmazott prior és likelihood függvény egy természetes konjugált párt alkot. E két feltételes eloszlás szórásának aránya határozza meg, hogy a prior és a likelihood milyen súllyal szerepeljen a posteriorban. Ezt az „egyensúlyozó" szerepet a fenti alakú modellben az m együttható játssza.

A posterior alakja ezek után meghatározható a prior és a likelihood függvény szorzataként:

$$\begin{gathered} f(w_{18},\dots,w_{65}\,|\,u_{18},\dots,u_{65}) \propto \\ \exp\left(-\sum_{x=18}^{64}(w_x-w_{x+1})^2\right)\exp\left(-\sum_{x=18}^{65} m_x(u_x-w_x)^2\right) = \\ \exp\left(-\sum_{x=18}^{64}(w_x-w_{x+1})^2-\sum_{x=18}^{65} m_x(u_x-w_x)^2\right). \end{gathered} \tag{19}$$

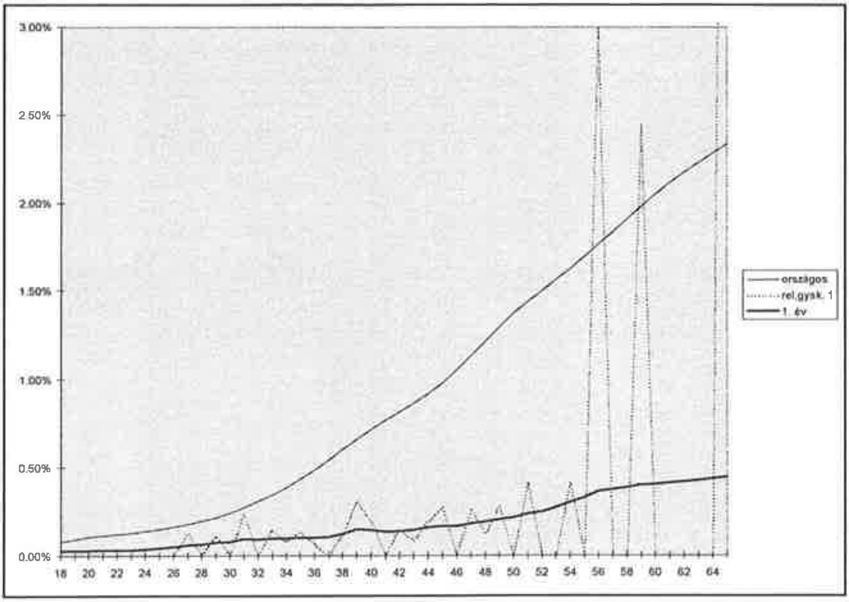
A halandósági ráták pontbecslésének tekintsük ezúttal az együttes posterior eloszlás várható értéke helyett a móduszát. Így nem szükséges meghatároznunk a fenti egyenletek arányossági tényezőit és az exponenciális függvény kedvező tulajdonságai miatt a számolás leegyszerűsödik. A módusz meghatározása a posterior sűrűségfüggvény maximumhelyének megkeresését jelenti. Ez ekvivalens a

$$\sum_{x=18}^{64}(w_x-w_{x+1})^2+\sum_{x=18}^{65} m_x(u_x-w_x)^2 \longrightarrow \min \tag{20}$$

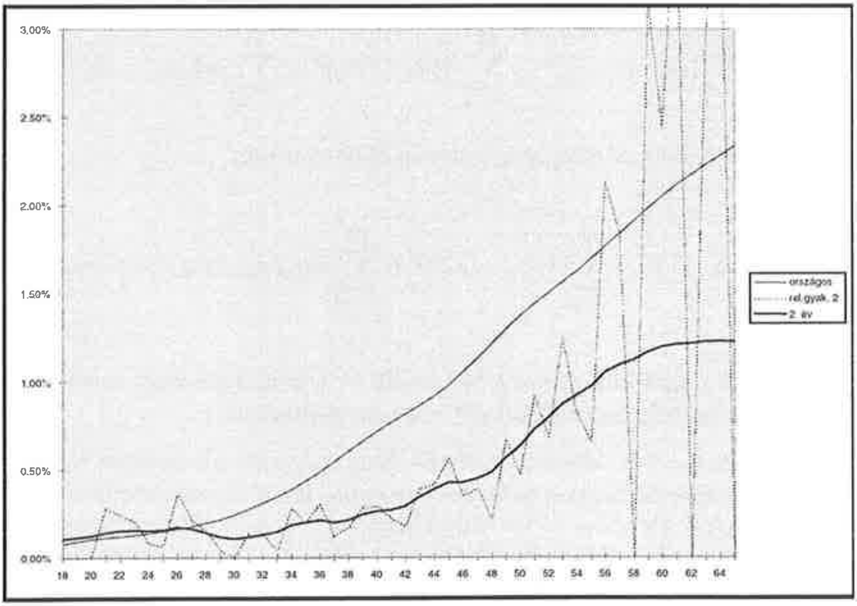
kvadratikus programozási feladat megoldásával, ami semmilyen elvi akadályba nem ütközik és számítógéppel a gyakorlatban is könnyen elvégezhető. A *4-6. ábrákon* a fenti módszerrel becsült halandósági görbe és relatív gyakoriságok láthatók, az 1., 2. és 3. év adatai, azaz rendre növekvő állományok alapján.

Mikor a második és harmadik év adatait is bevonjuk a modellbe, a priorban már érdemes szerepeltetni a korábbi becslések eredményét, mint előzetes feltételezést a halandósági ráták értékére. Legyen tehát a prior alakja a további években a következő:

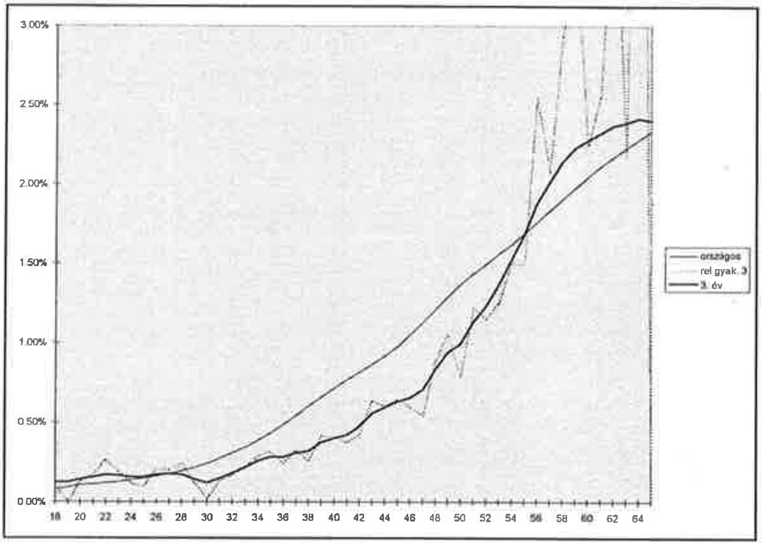
$$f(w_{18},\dots,w_{65}) \propto \exp\left(-\sum_{x=18}^{64}(w_x-w_{x+1})^2-\sum_{x=18}^{65} k(\hat{w}_x-w_x)^2\right). \tag{21}$$


3.00%
2.50%
2.00%
1.50%
1.00%
0.50%
0.00%
18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64
országos
rel.gyak. 1
1. év

4. ábra


3,00%
2.50%
2.00%
1.50%
1.00%
0.50%
0.00%
18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64
országos
rel.gyak. 2
2. év

5. ábra



6. ábra

A likelihood függvény nem változott, tehát a posterior az alábbi lesz:

$$f(w_{18},\dots,w_{65}\,|\,u_{18},\dots,u_{65}) \propto$$
$$\exp\left(-\sum_{x=18}^{64}(w_x-w_{x+1})^2-\sum_{x=18}^{65}k(\hat{w}_x-w_x)^2-\sum_{x=18}^{65}ml_x(u_x-w_x)^2\right). \tag{22}$$

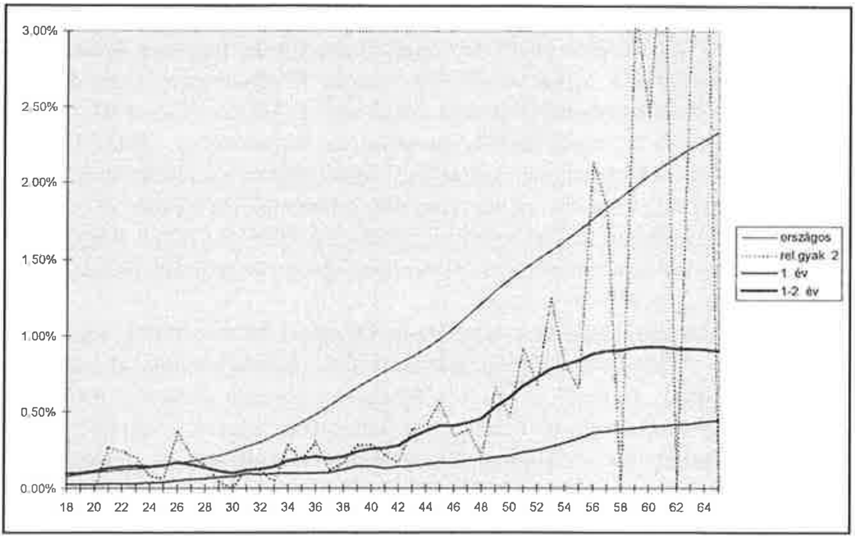
A megoldandó kvadratikus programozási feladat pedig:

$$\sum_{x=18}^{64}(w_x-w_{x+1})^2+\sum_{x=18}^{65}k(\hat{w}_x-w_x)^2+\sum_{x=18}^{65}ml_x(u_x-w_x)^2 \longrightarrow \min \tag{20}$$

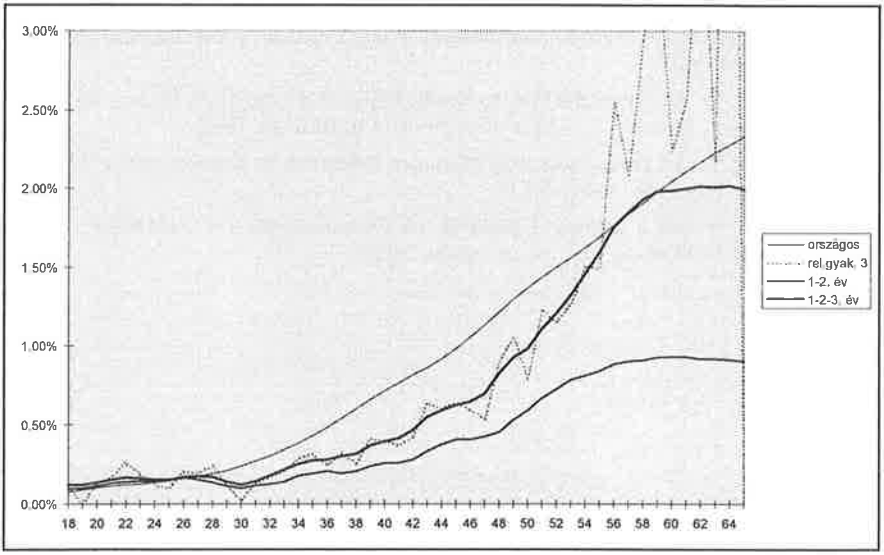
A *7-8. ábrák* a második illetve a harmadik év adatainak a fenti módon való bevonásával becsült halandósági görbéket tartalmazzák.[11]

A 4-6. illetve 7-8. ábrákon látható, hogy nagyobb állományok esetén a relatív gyakoriságok súlya a becslésben megnő, így a becslés szinte a megfigyelési adatokra illesztett görbe lesz. Megfigyelhető az is, hogy ez a priorban megfogalmazott simasági elvárás rovására mehet.

[11] m értékét 0.0001-nek, k-t 0.01-nek választottam.


3.00%
2.50%
2.00%
1.50%
1.00%
0.50%
0.00%
18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64
országos
rel.gyak. 2
1. év
1-2. év

7. ábra


3.00%
2.50%
2.00%
1.50%
1.00%
0.50%
0.00%
18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64
országos
rel.gyak. 3
1-2. év
1-2-3. év

8. ábra

3 Összefoglalás

E dolgozatban egy olyan problémára kerestem megoldást, melyre a bayesi elvek alapján való megközelítés kézenfekvőnek tűnik, hiszen a halandósági rátákkal kapcsolatban a biztosítók sok mintán kívüli információra támaszkodhatnak és ezek figyelembevétele a becsléskor a bayesi módszerek lényege. A külső információk közül kettőt, az országos halandósági táblát illetve a halandósági görbe simaságára vonatkozó feltételezést emeltük ki és ezekre alapoztunk egy-egy modellt. A két modellből hasonló, de természetesen nem megegyező eredmények származtak. Figyelemre érdemes, hogy szignifikáns, azonos irányú eltérést mutattak a biztosítási gyakorlatban általánosan használt halandósági táblához képest.

A módszerek hatékonysága nagy érzékenységet mutat a minta, azaz a rendelkezésre álló biztosítási állományok méretére. A magyarországi biztosítók tapasztalatai már a közeli jövőben elégségesek lesznek a bayesi módszerek eredményes használatára, az országosan összesített adatok alapján pedig — ha ilyenek léteznének— már jelenleg is érdemes lenne ilyen jellegű becsléseket készíteni.

Irodalom

1. Bayes, T.: An Essay Towards Solving a Problem in the Doctrine of Chances. The Philosophical Transactions, 1763. (Reprint in Biometrica 45 (1958))
2. Hunyadi L.: Bevezetés a maximum likelihood elméletbe. Oktatási segédlet. BKE Statisztika Tanszék, Budapest, 1992.
3. Hunyadi L.: A bayesi statisztika alapjai. Oktatási segédlet. BKE Statisztika Tanszék, Budapest, 1996.
4. Kovács E.: Kárstatisztikai elemzések. A BKE aktuárius főszakirány jegyzete. Budapest, 1997.
5. Lindley, D. V.: Introduction to Probability and Statistics from a Bayesian Viewpoint. Cambridge University Press, Cambridge, 1965.
6. Zellner, A.: An Introduction to Bayesian Inference in Econometrics. J. Wiley & Sons Inc., New York, 1971.
7. Zellner, A. (ed.): Bayesian Analysis in Econometrics and Statistics. North-Holland Publishing Co., Amsterdam, 1980