Tímarit um raunvísindi og stærðfræði
3. árg. 1. hefti 2005 – raust.is/2005/1/03

Spágildi skoðanakannana

Helgi Tómasson

Háskóla Íslands

Vefútgáfa: 30. desember 2005

Ágrip – Mælingum er safnað í tíma í þeim tilgangi að meta ákveðið ástand. Ástandið í þessu tilviki er það hlutfall þýðis sem svarar tiltekinni spurningu játandi. Gert er ráð fyrir að hlutfallið þróist með tímanum. Upplýsingar munu því úreldast ef mælingum er ekki safnað. Reglu Bayes má nota til að flétta saman gamlar upplýsingar við nýjar mælingar. Besta spá fæst með því að finna bestu leið til að vega saman eldri upplýsingar og nýjar mælingar. Sett er upp einfalt líkan og Bayes-tölfræði notuð til að rökstyðja val á vægi eldri upplýsinga. Sem sýnidæmi er aðferðinni beitt á niðurstöður skoðanakannana úr dagblaði.

1. Inngangur

Spá (e. prediction, forecasting) er ályktun um framtíð. Hugtakið spá lýsir því óvissu umfram nútíð, þ.e. stundin er ekki runnin upp og því ekki nokkur leið að komast að hinnu sanna. Þetta er frábrugðið mati (e. estimation) í venjulegri tölfræðilegu þar sem e.t.v. er fræðilega hægt að komast að hinnu sanna með því að mæla allt þýðið (e. population). Tölfræðilegu mati fylgir óvissa sem er háð stærð þýðis og úrtaks ásamt eiginleikum matsaðferðar. Matsaðferðir hafa ákveðinn breytileika (e. variance) og hlutdrægni (e. bias) Hlutdrægni getur verið af ýmsum uppruna. Skaðleg hlutdrægni er t.d. hlutdrægni sem ekki stefnir á núll þegar úrtak stefnir á allt þýðið. Oft má gera ráð fyrir að skaðleg hlutdrægni komi fyrir í hagnýtum rannsóknum vegna mæliskekkju eða vegna þess að mæliferlið (e. sampling process) er tengt því sem ályktа á um. Slík tengsl má t.d. vel ímynda sér í skoðanakönnunum þannig að já-hópur sé miklu viljugri að gefa upp skoðun en nei-hópur. Þegar slíkt kemur upp þarf tölfræðilegt líkan fyrir mæliferlið að vera samtvinnað líkani fyrir ferlið sem ályktа á um. Wooldridge (2002) gefur dæmi um slík líkön, sem nota má m.a. fyrir óslembin (e. non-random) úrtök. Ljóst er að í daglegu máli gera margir ekki greinarmun á spá og mati. Matsvandamál eru ekki viðfangsefni þessarar greinar. Viðfangsefnið hér er tímaraðagreining (e. time-series analysis) þar sem skoðuð eru tengsl úrtaksskekkju (e. sampling error) og þróunar í tíma. Tilgangur skoðanakannana er að gefa upplýsingar um ástand. Ef ástandið þróast í tíma skiptir tímapunkturinn sem könnunin fer fram á máli. Ef gert er ráð fyrir að framtíðin sé óvissari en nútíðin er eðlilegt að afskrifa gamlar kannanir. Hversu hratt sú afskrift á að gerast er háð því hversu hratt ástandið þróast í tíma.

Í þessari grein er ætlunin að setja fram einfalda túlkun á endurteknum skoðanakönnunum með formlegum hætti. Gögn úr skoðanakönnunum Fréttablaðsins eru notuð í tölulegt dæmi. Ástandið sem ályktа á um er fylgi við flokk og upplýsinga er aflað með spurningakönnunum.

Uppsetningin byggir á endurtekinni notkun á reglu Bayes. Upplýsingar eru settar fram í formi líkindadreifingar, gögnum er safnað og upplýsingar síðan uppfærðar með reglu Bayes. Þetta er síðan endurtekið. Vel þekkt vinnubrögð í þessum anda er Kalman-sían fyrir normal tímaraðalíkan. Kalmansíu fyrir normal tímaraðalíkan og hvernig mætti beita henni við gögn úr skoðanakönnunum er lýst í kafla 2. Normala líkanið er velþekkt og auðtúlkanlegt. Það getur hins vegar ekki verið rétt fyrir skoðanakannanir því mælingarnar eru jákvæðar heiltölur. Það getur hins vegar verið ágæt nálgun. Í kafla 3 er lýst líkindadreifingum sem tengjast skoðanakönnunum og tenglsum þeirra í gegnum reglu Bayes. Til að fá einfaldar formúlur er notast við samofnar fyrirframdreifingar (e. conjugate-prior distribution), sjá nánar í bókum um Bayes-aðferðir (Bernardo og Smith, 1994; Koop,

2003; Lancaster, 2004). Í kafla 4 er settur fram túlkunarrammi fyrir uppfærslu upplýsinga um einn valkost jafnóðum og gögn berast. Það að eldri upplýsingar séu nothæfar við spár er vel þekkt fyrir tímaraðir þar sem mælingar eru samfelldar. Alþekkt einfalt form er veldisjöfnun (e. exponential-weigthed-moving-average (EWMA)), þar sem ný mæling er vegin saman við eldri upplýsingar. Í þssari grein er rökstutt hvernig leiða má út veldisjöfnun sem bestu spá fyrir tímaröð af talningargögnum. Með bestu spá er hér átt við væntanlegt framtíðargildi. Lauslegur samanburður er gerður á ástandsformsnálgun (e. state-space) fyrir samfelldar tímaraðir (Harvey, 1989) og aðferð Grunwald, Hamza og Hyndman (1997). Lýst er lauslega hvernig reikna má sennileikafall (e. likelihood function) fyrir afskriftarstika upplýsinga. Að lokum er sýnt tölulegt dæmi þar sem aðferðunum er beitt á gögn úr skoðanakönnunum Fréttablaðsins.

2. Tímaraðalíkan á ástandsformi

Hentugt form til að lýsa hreyfimynstri í tímaraðagögnum er ástandsformið (e. state-space form). Venjuleg framsetning af líkani á ástandsformi er með mælijöfnu og ástandsjöfnu. Hér mætti t.d. hugsa sér að mælijafnan sé:

$$K(t) = n(t)p(t) + \varepsilon(t) \tag{1}$$

og ástandsjafnan

$$p(t) = p(t-1) + \xi(t) \tag{2}$$

Jafna (1) lýsir hve margir, $K(t)$, gefa upp ákveðna skoðun í $n(t)$ manna úrtaki á tíma t. Sanna hlutfallið með þessa tilteknu skoðun á tíma t er $p(t)$. Frávikið $\varepsilon(t)$ er munur á væntanlegum fjölda og mældum fjölda í úrtakskönnuninni. Sé um slembiúrtak að ræða er $K(t)|p(t)$ tvíliðudreift með dreifni $n(t)p(t)(1-p(t))$. Jafna (2) lýsir hreyfimynstri $p(t)$ í tíma. Hér er gert ráð fyrir að

$$E(p(t)|p(t-1)) = p(t-1),$$

þ.e. að væntanleg breyting á fylgi sé 0. Frávikið $\xi(t)$ er breyting á fylgi milli tímapunkta. Gert er ráð fyrir $\varepsilon(t)$ og $\xi(t)$ séu óháðar hendingar bæði innbyrðis og yfir tíma. Dreifni $\xi(t)$, $\tau^2 = V(\xi(t))$, er einfaldur mælikvarði á stöðugleika fylgis við skoðun. Það er ljóst að jafna (2) getur ekki verið sannleikurinn, því ekkert þvingar $p(t)$ til að vera á milli 0 og 1. Ef $p(t)$ er ekki nálægt 0 eða 1 og $n(t)$ sæmilega stórt kemur þetta lítið að sök. Kosturinn við að setja hreyfimynstrið fram eins og í jöfnu (2) er að stikinn τ verður mjög auðtúlkanlegur. Stikinn τ er staðalfrávik á breytingu $p(t)$ á einni tímaeiningu. Hugsanlegt væri að skrifa ástandið t.d. sem

$$z(t) = \log\left(\frac{p(t)}{1-p(t)}\right)$$

og nota síðan

$$p(t) = \frac{\exp(z(t))}{1+\exp(z(t))}$$

í mælijöfnunni. Það hefði í för með sér að aldrei kæmu fyrir óleyfileg gildi á $p(t)$ en á móti yrði staðalfrávik á breytingu ástands $z(t)$ á tímaeiningu illtúlkanlegri. Ef dreifing $\varepsilon(t)$ og $\xi(t)$ er nálguð með normaldreifingu gefa reiknireglurnar fyrir Kalmansíu einfalda leið til að reikna sennileikafallið. Það má síðan hámarka með tölulegum aðferðum með tilliti til τ og e.t.v. upphafsgildis, $p(t_0)$. Spáformúlan verður veldisjöfnun, sjá nánar í t.d. Harvey (1989).

3. Nokkrar líkindadreifingar og regla Bayes.

Reiknireglurnar fyrir Kalman-síu eru endurtekin notkun á reglu Bayes. Það má túlka þessa endurteknu notkun sem söfnun upplýsinga. Hefðbundin tölfræði túlkar líkur sem mælikvarða á tíðni. Í Bayestölfræði er leyft að nota líkur sem mælikvarða á vissu. Vissunni um tiltekinn stika er lýst með líkindadreifingu. Áður en tilraun er framkvæmd eru upplýsingar/vissa um óþekktan stika settar fram á formi líkindadreifingar. Hún nefnist fyrirframdreifingin fyrir óþekkta stikann (e. a priori distribution). Fyrirframdreifing, líkan og mælingar eru síðan flétttuð saman með reglu Bayes og þá fæst eftirádreifing (e. a posteriori distribution) fyrir óþekkta stikann. Kalman-sían samanstendur annars vegar af spájöfnum, þar sem upplýsingar á tíma $t-1$ eru notaðar til að spá útkomu á tíma t og hins vegar af uppfærslujöfnum þar sem regla Bayes er notuð til að endurskoða vissu í ljósi nýrrar mælingar.

Regla Bayes tengir líkur á atburðinum A gefinn atburðurinn B, $P(A|B)$ við $P(B|A)$ og líkurnar á A og B.

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

Ef þéttifall mælingar x er $f(x|\theta)$ og vissunni um stikann θ er lýst með þéttifallinu $f(\theta)$, þá segir regla

Bayes að

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta} \quad (3)$$

Fallið $f(\theta|x)$ í jöfnu (3) er þéttifall eftirádreifingar θ og túlkar vissu um stikann θ gefin gögnin x.

Ef stikinn τ í kaflanum hér á undan væri 0 ber að túlka það þannig að engin þróun (í tíma) eigi sér stað og því fáist sífellt betri upplýsingar um $p = p(t) = p(t-1)$ þegar gögnum er safnað. Það að hægt er að skrifa niður Kalman-síureglurnar er vegna þess að gengið er út frá að bæði mæling, þ.e. líkan, fyrirframdreifing og eftirádreifing séu normal. Sagt er að normal fyrirframdreifing sé samofin normal líkani. Ef fyrirframdreifing er samofin líkani, þá er eftirádreifing úr sömu fjölskyldu og fyrirframdreifing og ferlið er sagt vera *lokað við gagnasöfnun*. Ef $(K/n(t)|p(t))$ er normaldreift með meðaltal $p(t)$ og þekkt staðalfrávik σ, fyrirframdreifing fyrir $p(t)$ er normal með meðaltal $p(t-1)$ og staðalfrávik τ, þá er eftirádreifing $p(t) = p(t|K = k)$ normal með meðaltal

$$\frac{k}{n(t)}\frac{\tau^2}{\sigma^2+\tau^2} + p(t-1)\frac{\sigma^2}{\sigma^2+\tau^2}$$

og dreifnin verður

$$\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}$$

Tvíliðudreifingin $B(p,n)$ lýsir hendingu K sem getur tekið gildin $0, 1, \ldots, n$. Líkindamassafall $B(p,n)$ er

$$P(K=k) = \binom{n}{k} p^k(1-p)^{n-k}.$$

Í tvíliðulíkani, $K|p(t) \sim B(n(t), p(t))$, gildir

$$E(K|p(t)) = np(t), \quad (4)$$
$$V(K|p)(t) = n(t)p(t)(1-p(t)). \quad (5)$$

Jöfnu (5) má nota til að rökstyðja að góð ágiskun á σ^2 sé $p(t-1)(1-p(t-1))/n(t)$. Ljóst er að ef nákvæmni á að vera mikil, þ.e. vissa á að vera mikil þarf annað hvort τ að vera lítið eða $n(t)$ að vera stórt. Hér var notuð normal nálgun þ.e. að þéttifall fyrir mælingu sé

$$f(k/n(t)|p(t)) \propto \exp[-\frac{(k/n(t)-p(t))^2}{2\sigma^2}]$$

og að fyrirframþéttifall fyrir $p(t)$ sé

$$f(p(t)|p(t-1)) \propto \exp[-\frac{(p(t)-p(t-1))^2}{2\tau^2}]\,.$$

Með reglu Bayes fæst þéttifallið fyrir eftirádreifinguna

$$f(p(t)|k) = \int_{-\infty}^{\infty} f(k|p(t))f(p(t)|p(t-1))dp(t-1) \quad (6)$$

Það að mæling og fyrirframdreifning séu normal og σ þekkt gerir að hægt er að reikna heildið í jöfnu (6) á lokaðu formi. Ef hins vegar

$$p(t-1)(1-p(t-1))/n(t)$$

er sett inn í stað σ^2 þá verður heildið ekki eins viðráðanlegt.

Stærðirnar $p(t-1)$ og $p(t)$ eru hlutföll og taka því gildi á bilinu $[0,1]$. Normal-nálgunin á fyrirframdreifingunni úthlutar því ómögulegum gildum jákvæðum líkum. Sömuleiðis getur mælingin $K(t)$ aðeins tekið heiltölugildi og normal líkan fyrir $K(t)$ því einungis nálgun á sönnu dreifingunni.

Beta-dreifingin lýsir samfelldri hendingu sem getur tekið gildi á bilinu $(0,1)$. Þéttifall beta-dreifingar er

$$f(p) \propto p^{\theta_1-1}(1-p)^{\theta_2-1}, \quad \theta_1 > 0,\ \theta_2 > 0.$$

Sé fyrirframdreifing p í tvíliðudreifingu beta-dreifing, þá er eftirádreifingin einnig beta, þ.e. beta-dreifingin er lokuð við gagnasöfnun. Þetta má leiða út frá reglu Bayes, því af

$$f(k|p(t)) \propto p(t)^k(1-p(t))^{n(t)-k}, \quad (7)$$
$$f(p(t)|\theta_1,\theta_2) \propto p(t)^{\theta_1-1}(1-p(t))^{\theta_2-1}, \quad (8)$$

leiðir að

$$f(p(t)|k) \propto p(t)^{k+\theta_1-1}(1-p(t))^{n(t)-k+\theta_2-1}. \quad (9)$$

Hægt er að heilda út $p(t)$ úr jöfnu (7) og þá fæst óskilyrt dreifing K, beta-tvíliðudreifing. Líkindamassafall beta-tvíliðudreifingar er

$$f(k) = P(K=k) \propto \int_0^1 p^k(1-p)^{n-k}p^{\theta_1-1}(1-p)^{\theta_2-1}dp,$$

væntigildi hennar er

$$E(K) = n\frac{\theta_1}{\theta_2+\theta_2} \quad (10)$$

og dreifni

$$V(K) = n\frac{\theta_1\theta_2(\theta_1+\theta_2+n)}{(\theta_2+\theta_1)^2(\theta_1+\theta_2+1)}. \qquad (11)$$

Fjölliðudreifingin, $MN(\boldsymbol{p}, n)$, lýsir margvíðri hendingu $\boldsymbol{K} = (K_1, \cdots, K_m)$ þannig að $K_1 + \cdots + K_m = n$ og $\boldsymbol{p} = (p_1, \ldots, p_m)$ þannig að $p_1 + \cdots + p_m = 1$. Líkindamassafall $MN(\boldsymbol{p}, n)$ er

$$f(k_1, \ldots, k_m) \propto \prod_{i=1}^{m} p_i^{k_i}$$

Tvíliðudreifing er fjölliðudreifing með $m = 2$, $p_1 = p$ og $p_2 = (1-p)$. Dirichlet-dreifingin, $D(\boldsymbol{\theta})$, lýsir m-víðri hendingu, $\boldsymbol{p}$, þar sem $\sum_{i=1}^{m} p_i = 1$. Þéttifall hennar er

$$f(\boldsymbol{p}) \propto \prod_{i=1}^{m} p_i^{\theta_i - 1}.$$

Fjölliðudreifingin er margvíð útkvíkkun á tvíliðudreifingu og Dirichlet-dreifingin er margvíð útvíkkun á beta-dreifingu.

Ef p er beta-dreifð með stika θ_1 og θ_2 og $K|p$ er tvíliðudreift, $B(n, p)$, þá er óskilyrt dreifing K beta-tvíliðudreift. Samsvarandi margvíð útvíkkun er að ef $\boldsymbol{p}$ er Dirichlet-dreifð og $\boldsymbol{K}|\boldsymbol{p}$ fjölliðudreift þá er óskilyrt dreifing $\boldsymbol{K}$ fjölliðu-Dirichlet.

Samsvarandi niðurstöður fyrir fjölliðulíkan ef $\boldsymbol{K}|\boldsymbol{p} \sim MN(n, \boldsymbol{p})$ og $\boldsymbol{p} \sim D(\theta_1, \ldots, \theta_m)$ eru

$$E(K_i) = np_i, \qquad p_i = \frac{\theta_i}{\sum_{j=1}^{m}\theta_j}, \qquad (12)$$

$$V(K_i) = n\frac{n + \sum_{i=1}^{m}\theta_i}{1 + \sum_{i=1}^{m}\theta_i}p_i(1-p_i), \qquad (13)$$

$$\mathrm{Cov}(K_i, K_j) = -n\frac{n + \sum_{i=1}^{m}\theta_i}{1 + \sum_{i=1}^{m}\theta_i}p_ip_j. \qquad (14)$$

4. Uppfærsla upplýsinga

Kalman-sían er vel þekkt aðferðafræði við meðhöndlun normal tímaraða. Hér er lýst gróflega hvernig má útfæra sömu hugmynd, þ.e. endurtekna notkun á reglu Bayes, fyrir tímaraðir sem fást við endurteknar skoðanakannanir.

Gerum ráð fyrir fylgi við tiltekna skoðun sé $p(t_0)$, á upphafstímapunkti t_0. Vissunni um $p(t_0)$, þ.e. að upplýsingunum um fylgið er lýst þannig að $p(t_0)$ er betadreifð (tvívíð Dirichlet-dreifing). Væntanlegt gildi slíkrar beta-dreifingar er

$$E(p(t_0)) = \frac{\theta_1}{\theta_1+\theta_2} \qquad (15)$$

og dreifni

$$V(p(t_0)) = \frac{\theta_1\theta_2}{(\theta_1+\theta_2+1)(\theta_1+\theta_2)^2}. \qquad (16)$$

Hugsanleg spá á tíma t_0 væri $E(p(t_0))$ og hlutlægt mat á óvissu væri $V(p(t_0))$. Skilgreinum nú t_0 sem nútímann og t_1 sem punkt í framtíðinni. Segjum að við teljum væntanlegt gildi bestu spár vera

$$E(p(t_1)) = E(p(t_0))$$

þ.e. engin væntanleg breyting, en að framtíðin sé óvissari en nútíðin, þ.e.

$$V(p(t_1)) \geq V(p(t_0)).$$

Stikinn $p(t_1)$ er að sjálfsögðu óþekktur og því verður notast við ágiskun, þ.e. fyrirframmat á $p(t_1)$. Gert er ráð fyrir að á tíma t_0 sé unnt að tákna upplýsingar um $p(t_0)$ með beta-dreifingu, þ.e.

$$p(t_0) = p(t_0|t_0) \sim D(\theta_1(t_0), \theta_2(t_0)). \qquad (17)$$

Upplýsingar um ástand á tíma t_1 gefnar upplýsingar á tíma t_0 á sama hátt með

$$p(t_1|t_0) \sim D(\theta_1(t_1|t_0), \theta_2(t_1|t_0)) \qquad (18)$$

og upplýsingar að lokinni mælingu á tíma t_1 með

$$p(t_1|t_1) \sim D(\theta_1(t_1|t_1), \theta_2(t_1|t_1)). \qquad (19)$$

Dreifingin í jöfnu (17) lýsir þeim upplýsingum sem fyrir hendi eru á tíma t_0. Spáin um ástand á tíma t_1 er táknuð $p(t_1|t_0)$ í jöfnu (18) og dreifingin í jöfnu (19) lýsir þeim upplýsingum sem fyrir hendi eru eftir mælingu. Af stærðfræðilegum ástæðum, er valin samofin fyrirframdreifing, þ.e. gert er ráð fyrir að $p(t_1|t_0)$ sé beta-dreift með sama væntanlega gildi en meiri óvissu, þ.e. dreifnin er meiri. Þeim eiginleika að $E(p(t_0|t_0))$ sé óbreytt en $V(p(t_1|t_0)) > V(p(t_0|t_0))$ má ná, sbr. jöfnur (15) og (16), með því að nota

$$\theta_1(t_1|t_0) = \omega\theta_1(t_0|t_0) \qquad (20)$$

$$\theta_2(t_1|t_0) = \omega\theta_2(t_0|t_0) \qquad (21)$$

þar sem $\omega < 1$. Auðvelt er að sjá að

$$\frac{V(p(t_0|t_0))}{V(p(t_1|t_0))} = \frac{\omega(\theta_1(t_0|t_0) + \theta_2(t_0|t_0)) + 1}{\theta_1(t_0|t_0) + \theta_2(t_0|t_0) + 1}.$$

Stikinn ω er eins konar afskriftarstuðull á upplýsingar. Það að $\omega = 1$ túlkast sem að gamlar mælingar

séu jafngóðar og nýjar og $\omega > 1$ mætti túlka þannig að framtíðin væri vissari en nútíðin. Hér er notað að ef fyrirframvissan $p(t_1|t_0)$ er beta-dreifð og mælingin tvíliðudreifð, þá er eftirávissan einnig beta-dreifð. Það er vegna þess að beta-dreifingin er samofin fyrirframdreifing tvíliðulíkansins. Dreifing $p(t_1)$ gefnar upplýsingar á tíma t_1 er því

$$p(t_1) \sim D(\theta_1(t_1|t_1), \theta_2(t_1|t_1)).$$

Þetta má sjá með því að nota reglu Bayes. Regla Bayes sameinar upplýsingarnar sem fyrir hendi voru á tíma t_0 og upplýsingar úr mælingunni $k(t_1)$. Með því að setja inn í jöfnu (9) fæst

$$\begin{aligned} f(p(t_1)|K(t_1)) &= k(t_1), t_0) \\ &\propto p(t_1)^{k(t_1)}(1-p(t_1))^{n(t_1)-k(t_1)} \\ &\times p(t_1)^{\theta_1(t_1|t_0)-1}(1-p(t_1))^{\theta_2(t_1|t_0)-1}. \end{aligned}$$

Þetta er þéttifall fyrir beta-dreifingu,

$$D(\theta_1(t_1|t_1), \theta_2(t_1|t_1))$$

með

$$\theta_1(t_1|t_1) = \theta_1(t_1|t_0) + k(t_1) \quad \text{og} \tag{22}$$

$$\theta_2(t_1|t_1) = \theta_2(t_1|t_0) + n(t_1) - k(t_1) \tag{23}$$

Jöfnur (20) og (21) eru notaðar til að spá fyrir um framtíðarástand á tíma, t_1, gefnar upplýsingar um ástand á tíma t_0 og jöfnur (22) og (23) nota mælingu á tíma t_1 til að uppfæra vissu. Upplýsingasöfnunin úr jöfnum (20) til (23) skilgreina ítrunarferli sem tekið er saman í töflu 1.

Þetta ferli má nota við endurteknar skoðanakannanir á tímum $t_1, t_2, \ldots$, úrtaksstærðirnar eru $n(t_1), n(t_2), \ldots$ og fjöldi þeirra sem svarar játandi er $k(t_1), k(t_2), \ldots$. Eftir hverja könnun er fyrir hendi mat á ástandi (θ_1, θ_2). Afskriftahraði upplýsinga úr eldri könnunum ákvarðast af stuðlinum ω. Út kemur runa af $(\theta_1(t_i|t_i), \theta_2(t_i|t_i))$ sem nota má til að reikna dreifingu $p(t_i|t_i)$. Svipaðar aðferðir mætti nota til að reikna dreifingu $p(t|t_n)$, fyrir hvaða t sem er.

Tafla 1. Ítrunarferli í upplýsingasöfnun.

skref 0	vel $\theta_1(t_0\|t_0)$ og $\theta_2(t_0\|t_0)$
skref 1	reikna spá $\theta_1(t_1\|t_0)$ og $\theta_2(t_1\|t_0)$
skref 2	reikna $\theta_1(t_1\|t_1)$ og $\theta_2(t_1\|t_1)$
skref 3	set $t_0 = t_1$, skref 1-3 endurtekin

5. Veldisástandslíkan

Notkun Kalman-síu byggist á hreyfimynstri sem lýsa má með ástandsformi (e. state-space form). Uppfærsluformúlurnar í kafla 4 má fá með veldisástandslíkani (e. power-state-model, PSM). Grunnhugmyndin í PSM er að dreifingu framtíðargildis á stika (eða breytum), $\alpha(t_1)$, að gefnum nútíðarupplýsingum $I(t_0)$, megi lýsa með

$$f(\alpha(t_1)|I(t_0)) \propto f(\alpha(t_0)|I(t_0))^{\omega} \tag{24}$$

þar sem f er þéttifall stikans $\alpha(t)$. Hugmyndin er að dreifing framtíðargildis skuli hafa hágildi í sama punkti en að halar skuli vera þykkari en halar í dreifingu nútíðargildis. Það hefur verið sýnt að hér skiptir máli á hvaða formi stikinn er settur fram. Fyrir tvíliðulíkan, $K(t) \sim B(p(t), n(t))$, er heppilegt að nota

$$\alpha(t) = \log\left(\frac{p(t)}{1-p(t)}\right) \tag{25}$$

og vinna með tvíliðudreifinguna á forminu

$$\begin{aligned} &f(k(t)|\alpha(t)) \\ &\propto \exp[k(t)\alpha(t) - n(t)\log\{1+\exp(\alpha(t))\}] \end{aligned} \tag{26}$$

Samofin fyrirframdreifing fyrir stikann $\alpha(t_1)$ er

$$\begin{aligned} &f(\alpha(t_1)|I(t_0)) \\ &\quad\propto \exp[m(t_1|t_0)\sigma(t_1|t_0)\alpha(t_1) \\ &\quad\quad - \sigma(t_1|t_0)\log\{1+\exp(\alpha(t_1))\}]. \end{aligned} \tag{27}$$

Sjá Bernardo og Smith (1994). Þar sem $m(t_1|t_0)$ og $\sigma(t_1|t_0)$ eru stikar. Með reglu Bayes eru uplýsingunum $k(t_1)$ bætt við $I(t_0)$ og þá fæst

$$\begin{aligned} &f(\alpha(t_1)|k(t_1), I(t_0)) \\ &\quad\propto \exp[(k(t_1) + m(t_1|t_0)\sigma(t_1|t_0))\alpha(t_1) \\ &- (\sigma(t_1|t_0) + n(t_1))\log\{1+\exp(\alpha(t_1))\}]. \end{aligned} \tag{28}$$

Jafna (28) er á sama formi og jafna (27), því samofin fyrirframdreifing var notuð. Uppfærðu gildin $m(t_1|t_1)$ og $\sigma(t_1|t_1)$ fást með því að leysa

$$m(t_1|t_1)\sigma(t_1|t_1) = k(t_1) + m(t_1|t_0)\sigma(t_1|t_0) \tag{29}$$

$$\sigma(t_1|t_1) = (\sigma(t_1|t_0) + n(t_1)) \tag{30}$$

Ef sett er inn í (24) fæst að

$$m(t_1|t_0) = m(t_0|t_0)\sigma(t_1|t_0) = \omega\sigma(t_0|t_0). \tag{31}$$

Ef $t = t_0$ og $t_1 = t + 1$ fæst

$$\sigma(t+1|t) = \omega\sigma(t|t) = \omega(\sigma(t|t-1) + n(t)). \quad (32)$$

Jafna (32) er mismunanajafna. Ef úrtaksstærð er föst, $n(t) = n$, er stöðug lausn $\sigma = n\omega/(1-\omega)$. Á sama hátt sést að

$$\sigma(t+1|t)m(t+1|t) = \omega\sigma(t|t)m(t|t). \quad (33)$$

Ef við setjum $\sigma = n\omega/(1-\omega)$, þá fáum við

$$m(t+1|t) = (1-\omega)k(t)/n + \omega m(t|t-1). \quad (34)$$

Jafna (34) er venjulega uppfærslujafnan fyrir spár með veldisjöfnun. Þessi nálgun er byggð á Grunwald, Hamza og Hyndman (1997). Með þessu sést að PSM-framsetningin fyrir $\alpha(t)$ gefur óspáanleika, þ.e. að $E(k(t+1)|I(t)) = n(t)p(t)$. Í Harvey (1989) fæst það sama með því að krefjast óspáanleika. Nálgun Harvey (1989) byggir ekki á veldisvísisfjölskyldustikuninni og því verður útleiðsla á jöfnu (34) ekki eins gagnsæ. Athyglivert er að jafna (34) er hliðstæð því sem fæst fyrir normaldreifingu. Ef úrtaksstærðinni $n(t)$ er haldið fastri fæst ástand með stöðugum upplýsingum, þ.e. nákvæmni sem svarar til þess að úrtaksstærðin sé $n\omega/(1-\omega)$. Auðvelt er að reikna út úr jöfnum (33) og (34) þó að úrtaksstærðin sé breytileg. Formúlurnar fyrir heildarupplýsingarnar $\sigma(t+1|t)$ og spáð hlutfall $m(t+1|t)$ verða aðeins flóknari. Vert er benda á að $\sigma = \theta_1 + \theta_2$, þar sem θ_1 og θ_2 eru stikar úr beta-dreifingu sbr. 4. kafla og að $\omega = 1$ þýðir að $\sigma(t_k) = \sigma(t_0) + \sum_{i<k} n(t_i)$, þ.e. að $\omega = 1$ þýðir að leggja má saman eldri kannanir. Hliðstæðan í normala ástandslíkaninu er $V(\xi_t) = 0$ í jöfnu (2), þ.e. ástand fast yfir tíma.

6. Mat á ω

Til hagnýtra nota verður að gefa sér gildi á ω (eða τ). Hægt er að giska á gildi út frá hyggjuviti eða að reyna að nota gögn til að meta hversu mikið hald er í gömlum könnunum. Ein leið til að meta ω er að nota aðferð mesta sennileika (e. maximum-likelihood). Þá þarf að reikna sennileikafallið (e. likelihood function), $L(\omega)$. Fyrir liggja mælingar $(k(t_1), \ldots, k(t_r))$ á tímapunktum $(t_1, \ldots, t_r)$ og úrtaksstærðinar eru $(n(t_1), \ldots, n(t_r))$. Heppilegt er að skrifa líkindamassafallið á forminu

$$f(k(t_1), \ldots, k(t_r)) = f(k(t_1)) \prod_{i=2}^{r} f(k(t_i)|k(t_{i-1})),$$

Tafla 2. Mat á ω og τ.

	B	D	F	S	U	Ó
$\hat{\omega}$	0.21	0.15	0.23	0.29	0.39	0.04
$\hat{\tau}$	0.016	0.031	0.011	0.019	0.009	0.06

þar sem $f(k(t_i)|k(t_{i-1}))$ er skilyrta líkindamassafallið eða spádreifingin þegar tekið hefur verið tillit til þess að framtíðin er óvissari en nútíðin. Sennileikafallið er

$$L(\omega, \theta_1(t_0), \theta_2(t_0)) = f(k(t_1), \ldots, k(t_r)). \quad (35)$$

Þar sem $(\theta_1(t_0), \theta_2(t_0))$ eru upphafsgildi á θ_1 og θ_2. Þar sem gengið er út frá því að p sé beta-dreift með stikum θ_1 og θ_2, þá má reikna skilyrtu líkindadreifinguna með því að heilda yfir p

$$\begin{aligned} f(k(t_i)|k(t_{i-1})) &= \int_0^1 f(k(t_i), p|k(t_{i-1}))dp \\ &\propto \int_0^1 p^{k(t_i)+\theta_1-1}(1-p)^{n(t_i)-k(t_i)+\theta_2-1} \\ &\propto \Gamma(\theta_1 + k(t_i))\Gamma(\theta_2 + n(t_i) - k(t-i)). \end{aligned} \quad (36)$$

Þetta er líkindamassafall beta-tvíliðudreifingar. Á hverjum tímapunkti t_i er fyrir hendi spáð gildi á θ_1 og θ_2, $\theta_1(t_i|t_{i-1})$ og $\theta_2(t_i|t_{i-1})$. Þessir stikar eru reiknaðir með ítrunarferlinu úr töflu 1. Sennileikafallið L er síðan hámarkað með tilliti til $(\omega, \theta_1(t_0), \theta_2(t_0))$.

Útleiðsla fyrir margvíða tilfellið er hliðstæð en það leiðir til margliðu-Dirichlet-dreifingar í stað beta-tvíliðudreifingarinnar í jöfnu (36). Formúlan fyrir líkindamassafallið verður aðeins flóknari því að ekki er hægt að notast við gammafallið, Γ, heldur verður að skrifa „factorial" föll beint. Í báðum tilfellum er hætta á tölulegum vanda, þ.e. það þarf að reikna hliðstæðu við $\Gamma(x)$ fyrir mjög stórt x. Þess vegna þarf að grípa til einhverra nálgunaraðferða. Hugsanlegt væri að vinna með logra og nota Stirling nálgun fyrir $n!$, en hér var valin sú leið að nálga beta-tvíliðudreifingu og margliðu-Dirichlet-dreifingu með normal-dreifingum þar sem fræðilegu gildin fyrir væntanlegt gildi og dreifni voru fengin úr jöfnum (10-11) og (12-13).

7. Greining á gögnum

Aðferðafræðin var reynd á gögnum sem fengin voru úr Fréttablaðinu á tímabilinu janúar 2003 til júli 2004. Þetta er 31 könnun. Fyrri hluta 2003 eru oftast 7 dagar á milli kannana en eftir það eru kannanir gisnari. Að meðaltali eru 20.7 dagar á milli kannana. Í

töflu 2 er sýnt mat á ω og τ, sem byggir á einvíðri greiningu. Eðlilegt er að bera saman $\hat{\omega}$ og $\hat{\tau}$. Það að $\omega = 0.2$ ber að túlka þannig að ef könnun á tíma t_0 var gerð með 800 manna úrtaki þá er spáin, $p(t_1|t_0)$, túlkuð eins og 160 manna úrtak. Ef aftur er gerð 800 manna könnun er þeim 160 bætt við o.s.frv. Ef gerðar eru 800 manna kannanir reglulega er jafnvægisupplýsingamagnið (steady-state) samkvæmt formúlu (33) samsvarandi $0.2/0.8 \cdot 800 = 200$ manna úrtaki. Það að $\hat{\omega} = 0.04$ fyrir Ó (óákveðna) ber að túlka Ó skoðunina beri að afskrifa hratt. Gildin á $\hat{\tau}$ eru mat á staðalfráviki $\xi(t)$ í jöfnu (2), þ.e. breytingu á hlutfalli milli tímapunkta. Það að $\hat{\tau}$ sé 0.06 fyrir Ó ber að túlka að staðalfrávik á breytingu milli kannana sé 6%. Það að $\hat{\tau}$ sé 0.03 fyrir D ber að túlka sem að breytingar á fylgi D séu hægari, þ.e. færri prósent á tímaeiningu, en breytingar á fylgi Ó.

Vel má hugsa sér að afskrifa eigi upplýsingar úr síðustu könnun meira en ella ef langt er um liðið frá henni. Þetta má auðveldlega setja upp þannig að ω sé fall af tímanum sem liðinn er frá síðustu könnun. Hér var valin sú leið að skilgreina

$$\omega(\Delta_j) = \exp(-(\omega_0 + \omega_1 \Delta_j)) \qquad (37)$$

þar sem kannanir eru er framkvæmdar á tímum, $t_1, \ldots, t_n$, tíminn milli kannana $\Delta_j = t_j - t_{j-1}$. Stikana ω_0 og ω_1 má síðan meta með því að hámarka sennileikafallið í jöfnu (35).

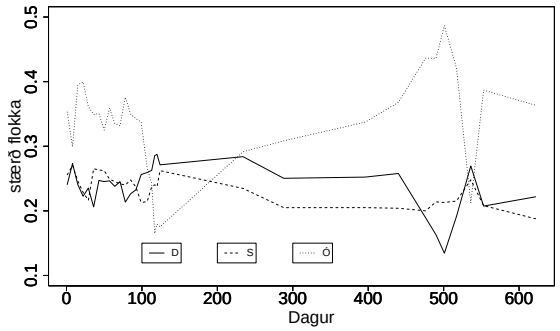
Í töflu 3 eru sýnd möt, $\hat{\omega}$, og staðalfrávik á stikamati, $s.e.(\hat{\omega})$, í formúlu (37) þar sem tími er mældur í mánuðum (mánuður=30 dagar). Fyrir alla flokka er fastinn ómarktækur og fyrir alla flokka nema U er hallatalan marktæk. Niðurstöður um þýðingu tíma milli kannana ber að taka með varúð þar sem mælingar eru fáar (31) og í mörgum tilfellum er tímabilið 7 dagar. Til að fá hugmynd um áhrif biðtímans þarf að hafa gögn með mislöngum tímabilum. Einnig er hugsanlegt að nota aðra framsetningu en í formúlu (37). T.d. mætti hugsa sér að nota aðeins tvö gildi, ω_s ef stutt er á milli og ω_l ef langt er á milli kannana. Mynd 1 sýnir þróun þriggja stærstu flokkanna yfir tíma. Óákveðnum fækkar rétt eftir dag 100. Dagur 1 var í janúar og kosningar framundan. Kosningar eru í maí, þ.e. fyrir dag 150. Eftir kosningar fjölgar óákveðnum aftur. Í kringum dag 500 fer D niður og Ó upp. Í kringum þann tímapunkt var í gangi umræða um lagafrumvarp varðandi eignarhald á fjölmiðlum. Mynd 2 sýnir metið fylgi við U-flokk fyrir 3 gildi á ω. Ef ω er sett jafnt og 0 þá er það hliðstætt því að henda gömlum könnunum, ef ω ser sett jafnt og 1 þá er það hliðstætt því að leggja saman allar kannanirnar. Í þriðja grafinu er nota mesta sennileika mat á ω. Samkvæmt töflu 2 er U-flokkurinn sá flokkur þar sem mest ástæða er til að taka tillit til liðinna kannana. Samkvæmt sömu töflu er Ó-flokkurinn sá flokkur þar sem minnst tillit ber að taka til liðinna kannana. Eins og sést á mynd 3 gefur metið ω mjög svipaða útkomu og það að setja $\omega = 0$. Sams konar mat má framkvæma fyrir margvítt líkan. Þá fékkst að $\hat{\omega} = 0.15$ en leitni í tíma var ekki marktækt frábrugðin núlli. Það að leitnistikinn í afskriftafallinu skuli vera mikið marktækur fyrir hvern flokk um sig en ekki marktækur fyrir margvíða líkanið þarf ekki að vera mótsögn. Óspáanleiki í mörgum víddum, þ.e. $E(\boldsymbol{p}(t+1|t)) = E(\boldsymbol{p}(t|t))$ setur ákveðinn skilyrði. Ef leyfa á mismunandi afskriftastika fyrir flokka í margvíðu líkani þýðir það að setja verður fram líkan á hreyfimynstri milli flokka. Höfundi, sem hefur fengist við margvíðar tímaraðir kemur ekki á óvart að það gangi illa að lýsa hreyfimynstri, t.d. þrívíðu röðinni í mynd 1 með einum stika, hvað þá öllum sex tímaröðunum. Ó-flokkurinn er í eðli sínu öðruvísi en hinir flokkarnir. Í gagnatímabilinu er kerfisbundinn skellur, þ.e. að kosningar koma fyrir á tímabilinu og rétt fyrir þær hrapar Ó-flokkurinn mikið. Gögnin byggjast á 25.925 svörum. Ef gert er ráð fyrir að 0.5% þýðisins sé spurt hverju sinni er eðlilegt að gera ráð fyrir að um það bil 1% svari þrisvar eða oftar og um það bil 15% tvisvar, þ.e. fjöldi einstaklinga á bak við þessi svör sé um það bil 22.000.

Tafla 3. Mat á afskrift upplýsinga í tíma, fyrir 5 flokka og óákveðna.

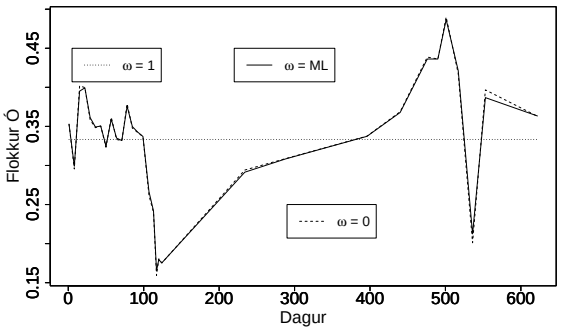
	B	D	F	S	U	Ó
$\hat{\omega}_0$	1.23	1.74	1.24	0.91	0.47	2.54
s.e.($\hat{\omega}_0$)	1.75	2.68	1.75	1.43	0.67	4.57
$\hat{\omega}_1$	0.54	0.13	0.23	0.38	0.48	0.67
s.e.($\hat{\omega}_1$)	0.13	0.03	0.06	0.15	0.47	0.10

8. Niðurlag

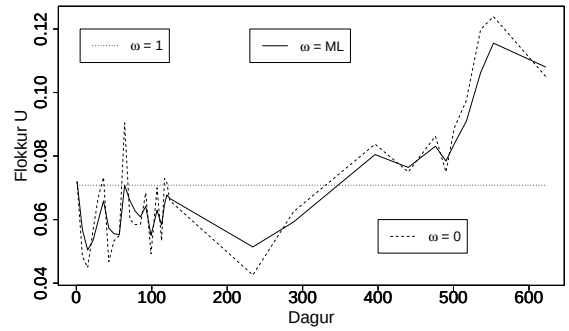
Aðferðafræðinni sem hér hefur verið lýst má flokka sem tímaraðagreiningu fyrir binomial(multinomial)-gögn. Hér var viðmiðunin sú að leitni væri ekki til staðar, þ.e. að spá skyldi vera jöfn mati á núverandi ástandi. Rökfræði bayesískrar tölfræði var beitt til að rökstyðja með hvaða hætti ætti að af-



Mynd 1. Þróun þriggja stærstu flokkanna.



Mynd 3. Metið fylgi Ó flokks fyrir 3 gildi á ω.



Mynd 2. Metið fylgi U flokks fyrir 3 gildi á ω.

skrifa gamlar upplýsingar. Sýnt var að það að ganga út frá leitnilausri spá var jafngilt ákveðnu ástandslíkani. Afskriftastikann má meta á hlutlægan hátt, t.d. með aðferð mesta sennileika. Einnig væri hugsanlegt að útfæra huglægt mat, t.d. með því að ákveða að upplýsingar skyldu helmingast á ákveðnum tíma. Skoðanakannanir eru alþekkt og vinsælt efni í fjölmiðlum og því voru valin þannig gögn hér til að gefa hugmynd um hvernig megi nota þessa aðferðafræði. Slíkt gagnasafn er ekki að öllu leyti heppilegt. Kannanirnar 31 úr Fréttablaðinu eru lítið gagnamagn í tímaraðagreiningu. Það þætti ekki gott að ætla að meta hreyfimynstur, t.d. með ARIMA greiningu með 30 mælingum.

Gagnatímabilið nær yfir kosningar og það gefur ferlinu ákveðinn skell. Í kringum kosningarnar gerðist það að fjölda óákveðinna fækkaði. Það er eðlilegt og ætti að byggja inn í líkanið. Það þýddi að einhvers konar form á leitni þarf að hanna. Það form verður háð undirliggjandi ferli, t.d. um pólitískar kannanir gæti gilt eitt form, um markaðskannanir gæti gilt annað. Fyrir þá sem nota kannanir reglulega er gott að hugleiða hvernig velja skuli stærð úrtaks, $n(t)$ og tímabilið á milli kannana, $t_j - t_{j-1}$. Ef $\omega = 1$ þá afskrifast gamlar upplýsingar ekki heldur hrúgast upp. Annars afskrifast gamlar kannanir og mat á afskriftafalli ω og liðnar úrtaksstærðir gefa til kynna upplýsingamagnið sem fyrir hendi er. Kostnaðurinn við að spyrja $n(t)$ manns ásamt afskriftafallinu ω ákvarða verð á upplýsingum. Fyrir þá sem ætla að safna upplýsingum og viðhalda þeim er hugsanlega skynsamlegt að byrja með lítil úrtök og dreifa þeim í tíma til að fá hugmynd um afskriftastikann ω. Ljóst er að virk upplýsingasöfnun þarf að vera í gangi til að hægt sé að meta $\omega(\Delta t)$ með nákvæmni.

Summary: Repeated opinion polls generate a sequence of binomial variables. The data are considered as noisy measurments of a proportion that evolves in time. At each point in time there is information available, which is a combination of older polls and a priori information. The information is updated by combining new data and prior information with Bayes rule. For simplicity conjugate priors and state-space representations are used. The optimal weighting of old and new information is estimated by numerically maximizing the likelihood function for the binary time series. Multivariate extensions are disscussed. The approach is illustrated by applying it to real data.

Heimildir

Bernardo, J. og Smith, A. (1994). *Bayesian Theory*. John Wiley & Sons Ltd.

Grunwald, G., Hamza, K., og Hyndman, R. (1997). Some properties and generaliztions of non-negative bayesian time series. *Journal of the Royal Statistical Society, B, 59*(3), 615–626.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Koop, G. (2003). *Bayesian Econometrics*. John Wiley & Sons. Ltd.

Lancaster, T. (2004). *An Introduction to Modern Bayesian Econometrics*. Blackwell Publishing.

Wooldridge, J.M. (2002). *Econometric Analysis of Cross Section and Panel Data* . MIT Press.

Um höfundinn: Helgi Tómasson lauk BS prófi i stærðfræði frá Háskóla Íslands 1977 og doktorsprófi í tölfræði frá Háskólanum í Gautaborg 1986. Með námi stundaði Helgi sumarvinnu í Seðlabanka Íslands, en starfaði einnig við kennslu, forritun og tölfræðiráðgjöf. Árið 1983 var Helgi við Purdue University í West-Lafayette í Bandaríkjunum. Árið 1985 vann Helgi við tölfræðideild IARC (alþjóðlegu krabbameinsrannsóknarstofnunarinnar) í Lyon í Frakklandi, 1986-1990 sem starfsmaður Kjararannsóknarnefndar og hefur frá 1990 verið fastráðinn kennari í tölfræði og hagrannsóknum við viðskipta- og hagfræðideild Háskóla Íslands. Rannsóknir Helga hafa einkum verið á sviði tölfræði fjármálamarkaða og tölfræði í læknisfræði.

Viðskipta- og hagfræðideild HÍ
Odda v/Sturlugötu
IS-101 Reykjavík
helgito@hi.is

Móttekin: 29. september 2004