Aktuarska matematika II, 2.dio

Bojan Basrak

2020

1. Bayesovska statistika

Apriori i aposteriori razdioba

Matematički gledano vjerojatnost je tek funkcija koja na matematički konzistentan način slučajnim dogadajima pridružuje brojeve izmedju 0 i 1, no kad govorimo o vjerojatnosti stvarnih dogadjaja, nju tek moramo interpretirati. Pošto je pitanje interpretacije vjerojatnosti od izuzetne važnost za gotovo sve znanosti nije neobično da su mu posvećene mnoge knjige i radovi. Od odgovora na ovo pitanje ovisi i naš pristup statističkom zaključivanju, a s obzirom na taj odgovor statistička teorija se ponekad dijeli na

- frekvencionističku
- Bayesovsku

◇ Osnovna ideja Bayesovske statistike je da parametri statističkih modela moraju i sami biti modelirani kao slučajni.
◇ Pri tome, mi i prije prikupljanja podataka imamo neku inicijalnu (više ili manje subjektivnu) ideju o razdiobi nepoznatog parametra.
◇ Inicijalna ili **apriori razdioba eng. prior** omogućuje nam da u model ugradimo npr. prethodno ili ekspertno znanje, no ona je izvor subjektivnosti u zaključivanju što je tipično i osnovni prigovor ovoj teoriji.
◇ Bayesovski pristup ipak dozvoljava da izračunamo vjerojatnosti oblika

$$P(\theta \in A)$$

gdje je A bilo koji skup u prostoru parametara, dok je u klasičnoj statistici parametar θ nepoznat, no fiksan, "bogomdan", pa ovakav izraz nema čak ni smisla računati.

Teorija nosi ime po svećeniku Thomasu Bayesu (1701-1761) i njegovom posmrtno objavljenom radu, a emocije koje je ovakav pristup izazivao medju klasičnim statističarima odlično ilustrira izjava jednog od njih, naime Maurice Kendall je primjetio:

> Life would be much simpler if the Bayesians followed the example of their master and published posthumuosly.

Bayesov rad bio je posvećen jednom od najelementarnijih problema statističkog zaključivanja, tj. procjeni vjerojatnosti uspjeha u slučajnom pokusu, ako nam je poznat ishod n nezavisnih ponavljanja ovakvog pokusa. Mi bismo kraće govorili o procjeni parametra p binomne slučajne varijable s parametrima n i p. Prilikom rješavanja ovog problema Bayes je korisitio i formulu ili teorem koji danas zovemo njegovim imenom. Naime ako sa H označimo proizvoljnu hipotezu o parametru slučajnog pokusa, a sa D njegov ishod tada je elementarno vidjeti da iz definicije uvjetne vjerojatnosti slijedi formula

$$P(H|D) = P(H)P(D|H)/P(D).$$

Izraz $P(H|D)$ označava vjerojatnost hipoteza nakon što smo vidjeli podatke i naziva se vjerojatnost **aposteriori, eng. posterior**. U slučaju da je naša hipoteza oblika $H = \{\theta = \theta_0\}$, tada kažemo da je hipoteza H jednostavna, a izraz $P(D|H)$ predstavlja uobičajenu vjerodostojnost. Vjerojatnost $P(H)$ ovisi o odabranoj apriori razdiobi za parametar θ, pa se ponekad ignorirajući nazivnik u gornjoj formuli ona piše kao

$$\text{aposteriori vjerojatnost} \propto \text{apriori vjerojatnost} \cdot \text{vjerodostojnost},$$

gdje $\propto$ stoji kao oznaka proporcionalnosti.

Ukoliko je θ npr. slučajna varijabla s neprekidnom razdiobom i gustoćom f tada je $P(H) = 0$ za sve jednostavne hipoteze. No gornju formulu možemo primjeniti i na uvjetne gustoće, tako da ako je naš uzorak $\mathbf{X}$ poprimio vrijednost $\mathbf{x}$, aposteriori razdiobu odredjuje aposteriori gustoća $f(\theta|\mathbf{x})$ za koju vrijedi

$$f(\theta|\mathbf{x}) = f(\theta)f(\mathbf{x}|\theta)/f(\mathbf{x})$$

ili ponovo

$$f(\theta|\mathbf{x}) \propto f(\theta)f(\mathbf{x}|\theta).$$

$\diamond$ Primjetite da aposteriori razdioba sadrži puno više informacija o parametru θ nego što je sadrži uobičajeni točkovni ili čak intervalni procjenitelji.

Primjer Prepostavimo prvo da je $X \sim B(n,p)$ broj uspjeha u n nezavisnih ponavljanja slučajnog pokusa. Neka je X poprimila vrijednost $k \in \{0,1,\ldots,n\}$. Po definiciji binomne razdiobe je

$$f(k|p) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Prepostavimo da je apriori razdioba parametra p beta razdioba s parametrima $a, b > 0$, tj. p ima apriori gustoću

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}.$$

Posebno je apriori očekivanje parametra p jednako $a/(a+b)$. Ako pretpostavimo da je $X|p \sim B(n,p)$. Aposteriori gustoća očito zadovoljava

$$f(p|k) \propto p^k (1-p)^{n-k} p^{a-1}(1-p)^{b-1} = p^{k+a-1}(1-p)^{n-k+b-1}.$$

Dakle aposteriori razdioba je ponovo beta s parametrima $k+a$ i $n-k+b$.

Za aposteriori očekivanje dobijemo

$$\frac{k+a}{k+a+n-k+b} = \frac{k+a}{n+a+b}.$$

Uočite da kako veličina uzorka n raste, za očekivati je da raste i broj uspjeha bez obzira na parametar $p \in (0,1)$, no tada je aposteriori očekivanje približno $\approx k/n$ što je uobičajeni procjenitelj za p izveden metodom maksimalne vjerodostojnosti npr.

$\diamond$ Ovo je karakteristično za Bayesovski pristup statističkom zaključivanju, naime kao što smo vidjeli u odredjivanju aposteriori razdiobe, ona je produkt apriori razdiobe i vjerodostojnosti koja ovisi o prikupljenim podacima, no kako broj podataka raste, raste i njihov utjecaj. Dok je za male uzorke aposteriori razdioba bitno odredjena upravo apriori razdiobom.

Napomena Kada apriori i aposteriori razdioba dolaze iz iste parametarske familije kao u gornjem primjeru kažemo da su konjugirane u odnosu na dani model. Takvi slučajevi su posebno zanimljivi jer je u njima tipično lakše pronaći aposteriori razdiobu, koja općenito može biti vrlo kompliciranа.

Funkcija gubitka

Ako bismo ipak htjeli izdvojiti jedan procjenitelj nepoznatog parametra Bayesovska statistika nas upućuje na korištenje **funkcije gubitka** L, koja usporedjuje našu procjenu sa stvarnom vrijednošću. Dakle prepostavimo ponovo da je sl. uzorak $\mathbf{X}$ poprimio vrijednost $\mathbf{x}$ te da nam je cilj naći veličinu $g(\mathbf{x})$ koja će aproksimirati nepoznati θ. Da bismo procjenili kvalitetu našeg procjenitelja možemo koristiti veličinu $L(g(\mathbf{x}), \theta)$.
Razumno je prepostaviti da je $L(\theta, \theta) = 0$, tj. gubitka nema ako procjenitelj pogadja vrijednost parametra. Jasno je da u netrivijalnim primjerima nije moguće naći procjenitelj koji pogadja θ egzaktno, no ono čemu se možemo realistično nadati je da je prosječni gubitak koji činimo u procjeni mali. Zato se kvaliteta procjenitelja zapravo ocjenjuje preko funkcije rizika, odn.

$$R(g(\mathbf{x})) = E[L(g(\mathbf{x}), \theta)] = \int L(g(\mathbf{x}), \theta) f(\theta|\mathbf{x}) \mathrm{d}\theta.$$

Naravno razne funkcije gubitka dat će nam i različite procjenitelje.

Najčešće korištena je **funkcija kvadratnog gubitka** tj.

$$L(g(\mathbf{x}), \theta) = (g(\mathbf{x}) - \theta)^2 .$$

Lako je vidjeti da rizik minimizira u ovom slučaju upravo **očekivanje** od θ u odn. na aposteriori razdiobu, pa je Bayesovski procjenitelj u ovom slučaju

$$E_{f(\cdot|\mathbf{x})}\theta = \int \theta f(\theta|\mathbf{x}) \mathrm{d}\theta$$

Drugi često korišten izbor je **funkcija apsolutnog gubitka**

$$L(g(\mathbf{x}), \theta) = |g(\mathbf{x}) - \theta|.$$

Ako prepostavimo da je θ realan parametar s vrijednostima u intervalu $(-\infty, \infty)$, te stavimo $g = g(\mathbf{x})$ za funkciju rizika vrijedi

$$R(g) = R(g(\mathbf{x})) = \int_{-\infty}^{g} (g - \theta) f(\theta, \mathbf{x}) \mathrm{d}\theta - \int_{g}^{\infty} (\theta - g) f(\theta, \mathbf{x}) \mathrm{d}\theta.$$

Koristeći pravila o deriviranju kompozicije funkcija nalazimo

$$\frac{\mathrm{d}}{\mathrm{d}g} R(g) = \int_{-\infty}^{g} f(\theta|\mathbf{x}) \mathrm{d}\theta - \int_{g}^{\infty} f(\theta|\mathbf{x}) \mathrm{d}\theta.$$

Tako da je kritična točka funkcije R upravo onaj g za koji je

$$\int_{-\infty}^{g} f(\theta|\mathbf{x}) \mathrm{d}\theta = \int_{g}^{\infty} f(\theta|\mathbf{x}) \mathrm{d}\theta \,,$$

tj. mora biti $P(g \geq \theta) = P(g \leq \theta)$. Odn. Bayesovski procjenitelj za funkciju aposlutnog gubitka je jednostavno **medijan** aposteriori razdiobe.

Kao treću alternativu možemo uzeti **funkciju gubitka sve ili ništa**

$$L(g(\mathbf{x}),\theta) = \begin{cases} 0 & \text{ako } g(\mathbf{x}) = \theta \\ 1 & \text{inače } . \end{cases}$$

Kako se ne možemo koristiti diferenciranjem da bismo pronašli minimum funkcije rizika u ovom slučaju, uvest ćemo aproksimativnu funkciju gubitka

$$L_\varepsilon(g(\mathbf{x}),\theta) = \begin{cases} 0 & \text{ako } |g(\mathbf{x}) - \theta| < \varepsilon \\ 1 & \text{inače } . \end{cases}$$

U slučaju da je aposteriori gustoća neprekidna funkcija, za male ε imamo aproksimaciju

$$R_\varepsilon(g) = EL_\varepsilon(g,\theta) = 1 - \int_{g-\varepsilon}^{g+\varepsilon} f(\theta|\mathbf{x})\mathrm{d}\theta \approx 1 - 2\varepsilon f(g|\mathbf{x})\,.$$

Pa se minimum od R_ε dostiže približno za maksimalne vrijednosti $f(g|\mathbf{x})$, a takve točke se zovu **mod** aposteriori razdiobe. Kako gubitak $L_\varepsilon \to L$ za $\varepsilon \to 0$, isto vrijedi i za minimum od R.

O apriori razdiobi

◇ Kako je proces statističkog zaključivanja u Bayesovskoj teoriji povezan s odlukom o apriori razdiobi, njen odabir je od ključne važnosti.
◇ U praksi se često koriste tzv. neinformativne apriori razdiobe. Takvom bismo npr. zvalu uniformnu razdiobu na intervalu $(0, 1)$ u primjeru procjene parametra p binomne razdiobe.
◇ Kako prethodna saznanja možemo na više načina ugraditi u apriori razdiobu, Bayesovska statistika neizbježno subjektivna u tom odabiru, što dakako stavlja povećanu odgovornost na statističara koji mora pravdati svoj odabir.

Neki od zahtjeva na apriori razdiobi su da mora svakako pokrivati sve uopće moguće vrijednosti parametra, te da ne smije biti previše lokalizirana. U praksi moramo voditi računa da je u odabiru apriori razdiobe lako napraviti grešku. Ovaj problem Bayesovci često ilustriraju riječima Olivera Cromwella pred saborom Škotske crkve iz 1650. godine:

> I beseech you, in the bowels of Christ, think it possible that you may be mistaken.

2. Teorija povjerenja

Uvjetna očekivanja

U ovom poglavlju bismo htjeli iskoristiti Bayesovski pristup statistici prilikom procjene razdiobe budućih šteta u nekom portfelju ili od nekog osiguranika. Odabrana razdioba će dakako utjecati i na visine premija, a njihovo odredjivanje jedan je od naših centralnih problema. Po onom što smo naučili do sada morat ćemo kombinirati do sada dostupne podatke o samom osiguraniku, sa našim prethodnim znanjem. Primjetite da historijskih podataka može biti vrlo malo, pomislite pri tom na osiguravatelje koji su procjenjivali rizik od urušavanja Twin Towers u New Yorku prije tragičnog napada ili na rizik kojem se danas izlažu osiguravatelji nekretnina u centru Los Angelesa npr. U prethodnom poglavlju, značajnu ulogu imale su uvjetne gustoće i uvjetne razdiobe, tako da ćemo prije nastavka ponoviti nekoliko glavnih rezultata vezanih i uz uvjetna očekivanja.

Sjetimo se da je $E(X|Y)$ zapravo slučajna varijabla koja "najbolje" opisuje sl. varijablu X kao funkciju od Y. Iz matematički rigorozne definicije slijedi

$$E(X) = E[E(X|Y)] \tag{2.1}$$
$$E(f(X)|X) = f(X) \tag{2.2}$$
$$E(Xf(Y)|Y) = f(Y)E(X|Y) \tag{2.3}$$
$$E[E(X|Y) - X]^2 = \inf_g E[g(Y) - X]^2 \tag{2.4}$$
$$\tag{2.5}$$

gdje nam ova zadnja relacija zapravo govori da je uvjetno očekivanje u srednjekvadratnom smislu optimalna procjena za sl. varijablu X u terminima od Y.
U odn. na uvjetnu razdiobu možemo definirati i uvjetnu nezavisnost, tada kažemo da su sl. varijable X_1 i X_2 uvjetno nezavisne za dano Y. Naravno tada će vrijediti i

$$E(X_1X_2|Y) = E(X_1|Y)E(X_2|Y)\,.$$

Uočite da uvjetno nezavisne slučajne varijable ne moraju biti i stvarno nezavisne, one su to intuitivno tek onda kada nam je vrijednost od Y poznata. Takodjer nezavisne sl. varijable uvjetovanjem mogu izgubiti to svojstvo.

Povjerenje

Da bismo ilustrirali način na koji kombiniranjem podataka i prethodnih znanja možemo odrediti očekivanu vrijednost šteta u nekom portfelju pogledajmo jedan jednostavan

Primjer

▷ osiguranik: lokalna bolnica sa 100 zaposlenika

▷ štete nastaju u slučaju tužbe pacijenata zbog pogrešaka u liječenju.

▷ podaci o zadnjih četiri godine govore da je godišnji trošak u ovakve svrhe bio u prosjeku 240 000 hrk.

▷ iz velikog skupa nama dostupnih podataka o preostalim bolnicama u zemlji moguće procijeniti da je prosječna godišnja šteta za bolnicu od 100 zaposlenika 400 000 hrk.

No neke od preostalih bolnica posluju pod bitno drugačijim okolnostima, te su izložene većem riziku tretirajući kompleksnije slučajeve ili obradjujući nešto veći broj pacijenata po zaposleniku.

Naš zadatak je prije svega odrediti neto premiju za nadolazeću godinu za bolnicu koja tek pristupa osiguranju.

Odmah možemo ponuditi dva odgovora.
▷ Jedan je da iz podataka za dotičnu bolnicu zaključimo da je odgovarajuća neto premija 240 000, što će osiguravatelj dakako uvećati doplatkom za sigurnost i svojim administrativnim troškovima.
▷ Druga je mogućnost da zanemarimo relativno mali skup podataka o toj bolnici i neto premiju postavimo na 400 000, do kojih smo došli koristeći veći skup podataka, no oni su dakako nešto manje relevatni za tu konkretnu bolnicu.

Veća premija znači dakako i veći očekivani dobitak za osiguravatelja, no prevelika premija bi mogla ravnatelja bolnice otjerati k drugom osiguravatelju.

Za neto premiju mogli bismo odrediti i bilo koju konveksnu kombinaciju ova dva iznosa tj.

$$z \cdot 240\,000 + (1 - z) \cdot 400\,000$$

gdje je $z \in [0, 1]$ takozvani **faktor povjerenja**. Upravo ovaj odgovor dobijamo slijedimo li princip povjerenja. Iz izraza za premiju je jasno da će ona biti najbliža empirijskoj vrijednosti očekivanja što je z bliže 1. Ako z pada u ovom konkretnom primjeru premija će rasti k iznosu 400 000.
U principu, ako ocjenimo da je skup podataka koji nam je dao 400 000 kao očekivanu vrijednost štete nereprezentativan, ili ako se produlji period u kojem smo skupljali podatke povisivali bismo faktor z. No vrijedi i obrnuto, ako je duljina perioda za koji imamo podatke kraća, ili ako smo raspolagali podacima o velikom broju bolnica sličnih karakteristika, za z bismo uzeli nešto manju vrijednost.

Sažeto govoreći, neto premija u teoriji povjerenja konveksna je kombinacija oblika

$$z \cdot \bar{X} + (1 - z) \cdot \mu \tag{2.6}$$

gdje je $\bar{X}$ prosječna vrijednost prikupljenog uzorka, a μ je očekivana vrijednost šteta prema našim prethodnim saznanjima. Općenito će z će ovisiti o duljini našeg uzorka, kao i o stupnju uvjerenja koji imamo u očekivanu vrijednost šteta dobivenu iz apriori znanja. S druge strane ne želimo da faktor z ovisi o stvarnim vrijednostima ova dva iznosa. Naš cilj je koherentna metoda njegovog odredjivanja.

Bayesovsko povjerenje

Kako smo vidjeli Bayesovska statistika nudi prirodnu metodu kojom možemo kombinirati prethodno saznanje ili vrlo veliki skup podataka izražen u vidu apriori razdiobe s empirijskim podacima. Stoga ne iznenadjuje da je možemo iskoristiti kako bismo došli do faktora povjerenja u nekim slučajevima. Pretpostavimo kvadratnu funkciju gubitka, tako da je naš procjenitelj očekivanja šteta X, zapravo očekivanje od X u odnosu na aposteriori razdiobu. Detaljno ćemo obraditi dva važna modela.

Primjer (P/Γ model) Prepostavimo da sve štete u portfelju iznose točno 1. Tako da je ukupna šteta u nekom periodu odredjena jedino njihovim brojem, Prepostavimo nadalje da broj šteta u svakoj godini ima Poissonovu razdiobu s parametrom λ. Prepostavimo da parametar λ nije poznat, no na osnovu prethodnog iskustva i drugih podataka utvrdjeno da je λ približno ima $\Gamma(a, b)$ razdiobu.
Nama su nadalje, dostupni podaci o posljednjih n godina, koje modeliramo kao $P(\lambda)$ distribuirane sl. varijable $X_1, \ldots, X_n$ koje su uvjetno na λ takodjer i medjusobno nezavisne. Preciznije, naša prepostavka je da su $X_i|\lambda \sim P(\lambda)$ te n.j.d. Prepostavimo da je sl. uzorak

$$\mathbf{X} = (X_1, \ldots, X_n)$$

poprimio vrijednost

$$\mathbf{x} = (x_1, \ldots, x_n).$$

Nas zanima odredjivanje neto premija na osnovu ovih saznanja. Uočite prvo

$$E(X_{n+1}|\mathbf{X}) = E[E(X_{n+1}|\lambda, \mathbf{X})|\mathbf{X}] = E(\lambda|\mathbf{X})\,.$$

Označimo sa $X = X_{n+1}$ broj šteta u nadolazećoj godini. Kako je $E(X|\lambda) = \lambda$, prirodno je iskoristiti mogućnost da izračunamo Bayesovski procjenitelj nepoznatog parametra λ i njega koristimo kao procjenu neto premija. Imamo dakle tipičan problem Bayesovske statistike.
Iz iznesenih podataka slijedi da je apriori gustoća

$$f(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda),\ \lambda > 0,$$

a vjerodostojnost iznosi

$$f(\mathbf{x}|\lambda) = \frac{\lambda^{x_1+\cdots+x_n}}{x_1! \cdots x_n!} e^{-\lambda n}.$$

Tako da za aposteriori gustoću vrijedi

$$f(\lambda|\mathbf{x}) \propto \lambda^{x_1+\cdots+x_n+a-1} e^{-(b+n)\lambda}.$$

Možemo zaključiti da je aposteriori razdioba $\Gamma(x_1 + \cdots + x_n + a, b + n)$, stoga je Bayesovski procjenitelj za neto premiju

$$E[\lambda|\mathbf{X} = \mathbf{x}] = \frac{x_1 + \cdots + x_n + a}{b + n},$$

ako $\mathbf{X} = \mathbf{x}$. Primjetite da je apriori očekivanje od λ (pa dakle i od X) jednako a/b, dok bi uobičajeni empirijski procjenitelj tog očekivanja iznosio $(x_1 + \cdots + x_n)/n$. Ako postavimo $z = n/(b + n)$ uočite da vrijedi

$$E[\lambda|\mathbf{X} = \mathbf{x}] = z\frac{x_1 + \cdots + x_n}{n} + (1 - z)\frac{a}{b}.$$

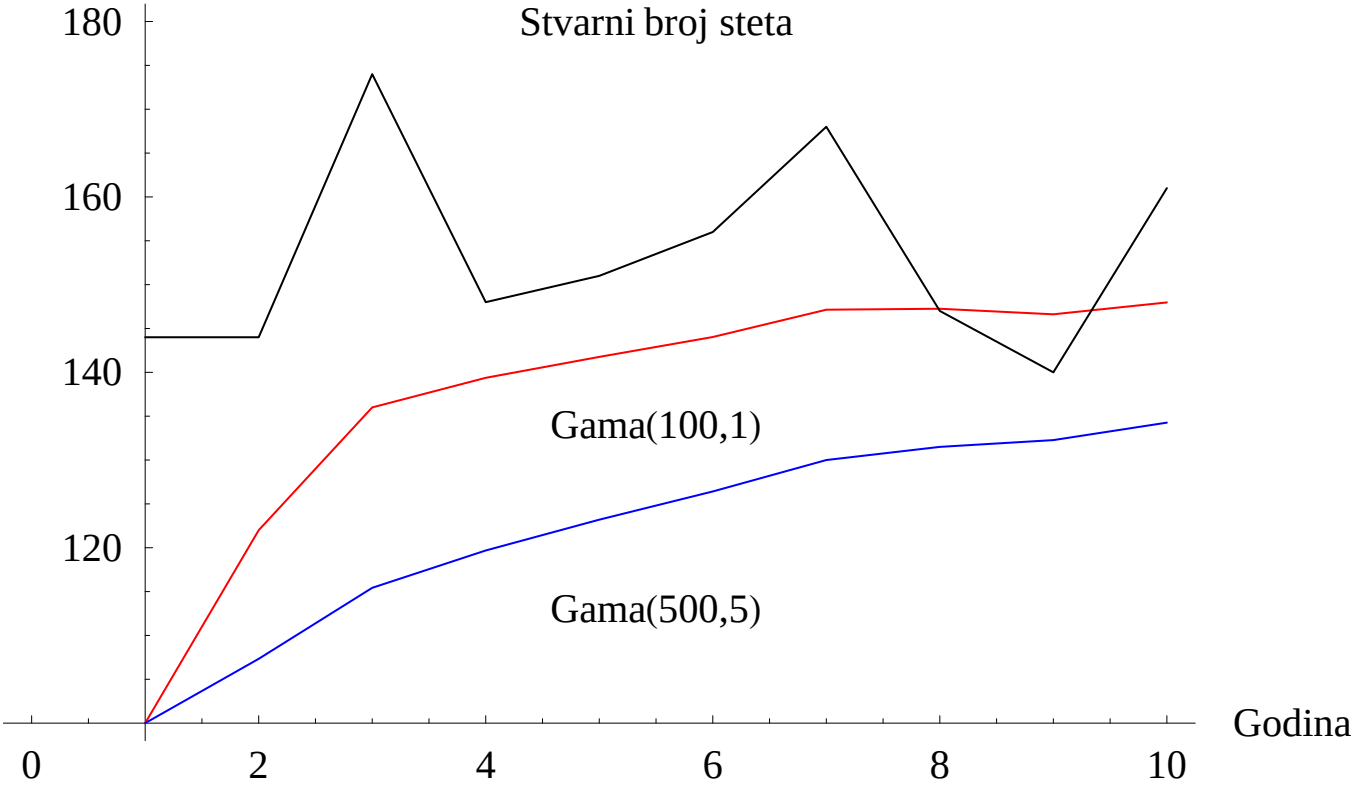
Dakle Bayesovski procjenitelj neto premije je konveksna kombinacija poput one koje smo sreli u teoriji povjerenja i izrazu (2.6). Ponovo je za velike n ovaj procjenitelj asimptotski jednak empirijskoj procjeni tj. aritmetičkoj sredini uzorka. Dok za male uzorke on značajno ovisi o apriori razdiobi. Posebno je u odsustvu svarnih podataka (tj. za $n = 0$), Bayesovska procjena jednaka a/b, tj. očekivanju u odn. na apriori razdiobu. Primjetite nadalje da faktor povjerenja pada k 0 kako raste parametar b, tj. kako se smanjuje varijanca apriori razdiobe i s njom neizvjesnost o stvarnoj vrijednosti parametra.

U knjizi je ovaj model ilustriran na jednom jednostavnom primjeru:
◇ u 10 sukcesivnih godina broj šteta ima sljedeće vrijednosti:

$$144, 144, 174, 148, 151, 156, 168, 147, 140, 161.$$

◇ Prepostavimo da ovi brojevi dolaze iz P/Γ modela uz dvije apriori razdiobe na parametar λ (poznato je $\lambda = 150$ jer su brojevi simulirani)

$$\Gamma(500, 5) \quad \text{i} \quad \Gamma(100, 1)$$



Primjetite da iako obje apriori razdiobe imaju očekivanje 100, razdioba $\Gamma(500, 5)$ *ima bitno manju varijancu, što uzrokuje sporiju konvergenciju Bayesovskog procjenitelja k stvarnoj vrijednosti.*

Primjer (N/N model) Za razliku od prethodnog primjera prepostavite da veličine $X_1, \ldots, X_n$ označavaju ukupne štete u n minulih godina, naš cilj je ponovo odrediti očekivani iznos štete $X = X_{n+1}$ u narednoj godini. Prepostavimo da je uzorak $(X_1, \ldots, X_n)$ poprimio vrijednost $\mathbf{x} = (x_1, \ldots, x_n)$. Ovoga puta prepostavimo da štete slijede normalnu razdiobu $N(\theta, \sigma_1^2)$ uz iste oznake. Prepostavimo da nam je parametar σ_1 poznat, dok za parametar θ možemo samo prepostaviti da i sam dolazi iz apriori razdiobe $N(\mu, \sigma_2^2)$, gdje su μ i σ_2 poznati parametri. Prema tome, uvjetovano na θ opažene štete $X_1, \ldots, X_n$ kao i šteta X u narednoj godini su nezavisne i distribuirane kao $N(\theta, \sigma_1^2)$.
I ovdje je

$$E(X_{n+1}|\mathbf{X}) = E[E(X_{n+1}|\theta, \mathbf{X})|\mathbf{X}] = E(\theta|\mathbf{X})\,.$$

Jasno je da je apriori gustoća parametra θ

$$f(\theta) \propto \exp\left(\frac{-(\theta-\mu)^2}{2\sigma_2^2}\right),\ \theta \in \mathbb{R},$$

a vjerodostojnost iznosi

$$f(\mathbf{x}|\theta) \propto \exp\left(\frac{-\sum_{i=1}^n (x_i-\theta)^2}{2\sigma_1^2}\right),\ \mathbf{x} \in \mathbb{R}^n,$$

tako da se lako vidi kako aposteriori gustoća zadovoljava

$$f(\theta|\mathbf{x}) \propto \exp\left(\frac{-\sum_{i=1}^n (x_i-\theta)^2}{2\sigma_1^2}\frac{-(\theta-\mu)^2}{2\sigma_2^2}\right),\ \theta \in \mathbb{R}\,.$$

Sredjivanjem gornjeg izraza po θ, slijedi da je aposteriori razdioba od θ uz dano $\mathbf{X} = \mathbf{x}$

$$N\left(\frac{\sigma_1^2\mu + \sigma_2^2 n\bar{x}}{\sigma_1^2 + n\sigma_2^2}, \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + n\sigma_2^2}\right),$$

gdje je $\bar{x} = (x_1 + \cdots + x_n)/n$.

Tako da je pod uvjetom $\mathbf{X} = \mathbf{x}$

$$E[\theta|\mathbf{X}] = \frac{\sigma_1^2\mu + \sigma_2^2 n\bar{x}}{\sigma_1^2 + n\sigma_2^2} = \frac{\sigma_1^2}{\sigma_1^2 + n\sigma_2^2}\mu + \frac{\sigma_2^2 n}{\sigma_1^2 + n\sigma_2^2}\bar{x}.$$

Dakle, ponovo dobijamo izraz oblika (2.6) uz

$$z = \frac{n}{n + \sigma_1^2/\sigma_2^2}.$$

Pa je i u ovom modelu Bayesovski procjenitelj moguće zapisati kao konveksnu kombinaciju oblika (2.6) uz dobro odabrani faktor povjerenja. Kao i prije faktor povjerenja raste k 1, za $n \to \infty$, te pada k 0, ako se σ_2 smanjuje (i opet s njim se smanjuje i neizvjesnost o stvarnoj vrijednosti parametra θ).

Primjetite da su **osnovne prepostavke** oba ova Bayesovska modela

- svaku pojedinu policu opisuje par $(\theta, (X_i))$, gdje je sl. varijabla θ tzv. **parametar rizika ili heterogenosti**, a niz sl. varijabli (X_i) predstavlja niz šteta u danoj polici.,
- razdioba od X_i ovisi o slučajnom parametru θ,
- uz dano θ sl. varijable X_i su n.j.d..

U gornjim primjerima bila nam je poznata i razdioba parametra θ, tj. apriori razdioba.
Nadalje primjetite da model ne implicira da su X_i i bezuvjetno medjusobno nezavisne,. Promotrimo npr. N/N model ponovo, iz prepostavki slijedi

$$E(X_1X_2) = E[E(X_1X_2|\theta)] = E[E(X_1|\theta)E(X_2|\theta)] = E[\theta^2] = \sigma_2^2 + \mu^2.$$

Za nezavisne sl. varijable ovo očekivanje moralo bi biti jednako $E(X_1)E(X_2)$. S druge strane imamo

$$E(X_1) = E(X_2) = E(E(X_2|\theta)) = E(\theta) = \mu\,.$$

Dakle nezavisnost slijedi jedino u slučaju da je $\sigma_2 = 0$, tj. samo kada parametar θ nije slučajan već unaprijed poznat.

Situacija koju smo vidjeli u prethodna dva primjera nažalost nije uvijek ovako dobra. Bayesovski procjenitelj općenito ne mora biti linearna funkcija sl. uzorka $\mathbf{X}$, pa se općenito ne može zapisati u obliku (2.6) koristeći faktor povjerenja. Štoviše Bayesovski procjenitelj općenito ne možemo egzaktno izračunati.
Dodatni problem u praksi je i kako doći do ostalih parametara modela, npr. u slučaju P/Γ modela morali bismo znati parametar b prije statističke analize. Uočite da nismo diskutirali kako odabrati te parametre. U praksi je jasno da će oni odražavati naša subjektivna uvjerenja o stvarnim vrijednostima parametra θ. Uzgred primjetimo da θ u Bayesovskim modelima može općenito biti i višedimenzionalan parametar, dakle ne nužno realna sl. varijabla.

Empirijska Bayesovska teorija povjerenja (model 1)

Preuzet ćemo i dalje osnovne pretpostavke Bayesovskih modela, no ovoga puta nećemo prepostaviti da nam je poznata apriori razdioba parametra θ. Preciznije naše prepostavke su

- svaku pojedinu policu (ili dio portfelja) opisuje par $(\theta, (X_i))$, gdje je parametar θ nepoznat, a niz sl. varijabli (X_i) predstavlja niz šteta u danoj polici.,
- razdioba od X_i ovisi o slučajnom parametru θ,
- uz dano θ sl. varijable X_i su n.j.d..

Cilj nam je ponovo odrediti neto premiju, odn. očekivanu vrijednosti od $X = X_{n+1}$ ako nam je dan sl. uzorak $\mathbf{X} = (X_1, \ldots, X_n)$.

Uvedimo sljedeće oznake

$$m(\theta) = E(X_i|\theta) \quad \text{i} \quad s^2(\theta) = \mathrm{var}(X_i|\theta)\,.$$

Kako su uz dano θ sl. varijable X_i jednako distribuirane, $m(\theta)$ i $s^2(\theta)$ ne ovisi o indeksu i. Primjetite nadalje da su, s obzirom na to da je θ sl. varijabla i ove dvije vrijednosti takodjer sl. varijable. Ukoliko bi nam θ bio poznat, $m(\theta)$ bi upravo bio tražena neto premija, no naš cilj je kao i do sada procijeniti $m(\theta)$ uz dano $\mathbf{X}$. Bayesovski procjenitelj koji smo do sada koristili bio je

$$E[m(\theta)|\mathbf{X}]\,,$$

i davao je jasan odgovor na traženo pitanje. No on nam kao što smo već istakli može biti vrlo komplicirana, a ne nužno linearna funkcija našeg uzorka i poznatih veličina.

Dodatne teškoće s kojima se dakle susrećemo su

- $m(\theta)$ nije nužno oblika $m(\theta) = \theta$ kao u prethodna dva modela,
- $s^2(\theta)$ nije nužno konstanta kao u N/N modelu,
- parametar θ nije nužno realan broj,
- apriori razdioba od θ nam nije poznata.

U novim i težim okolnostima pokušat ćemo umjesto Bayesovskog procjenitelja, naći samo optimal procjenitelj za $m(\theta)$ medju linearnim funkcijama našeg uzorka tj. medju procjeniteljima koji se mogu zapisati kao

$$a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n \, .$$

Kao kriterij optimalnosti ovog linearnog Bayesovskog procjenitelja i dalje možemo koristiti kvadratnu funkciju gubitka, pa je naš cilj zapravo pronaći konstante $a_0, a_1, \ldots, a_n$ tako da minimiziramo sljedeći izraz

$$E\left(m(\theta) - (a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n)\right)^2.$$

Srednjekvadratna pogreška glatka je funkcija parametara a_i. Zato njen minimum možemo tražiti diferenciranjem. Parcijalna derivacija po a_0 nam daje sljedeću jednadžbu u točki minimuma

$$E\left(m(\theta) - (a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n)\right) = 0.$$

Pa je prema definiciji funkcije m, iz $EX_i = E(E(X_i|\theta)) = Em(\theta)$

$$a_0 = E\left(m(\theta)\right)\left(1 - \sum_{i=1}^{n} a_i\right). \tag{2.7}$$

Ako diferenciramo srednjekvadratnu pogrešku po a_i, $i > 0$, dobijemo uvjet

$$E\left(X_i[m(\theta) - (a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n)]\right) = 0. \tag{2.8}$$

Sad uočimo da vrijedi

$$E(X_i m(\theta)) = E[E(X_i m(\theta)|\theta] = E(m(\theta)^2) = \mathrm{var}(m(\theta)) + [E(m(\theta))]^2 .$$

Slično je i za $j \neq k$

$$E(X_j X_k) = \mathrm{var}(m(\theta)) + [E(m(\theta))]^2 .$$

Na kraju uočite

$$\begin{aligned} E(X_i^2) &= E[E(X_i^2|\theta)] = E[s^2(\theta)] + E[m^2(\theta)] \\ &= E[s^2(\theta)] + \mathrm{var}[m(\theta)] + (E[m(\theta)])^2. \end{aligned}$$

Tako da se (2.8) može zapisati i kao

$$a_i E[s^2(\theta)] = \left(1 - \sum_{i=1}^{n} a_i\right) \left(\mathrm{var}[m(\theta)] + (E[m(\theta)])^2 - a_0 E[m(\theta)]\right). \tag{2.9}$$

Odavde je jasno (mada smo to mogli zaključiti i zbog simetrije) da za optimalne parametre mora biti

$$a_1 = a_2 = \ldots = a_m.$$

Označimo $z = \sum_{i=1}^{n} a_i$, tj. $a_i = z/n$, za $i \geq 1$. Tako da se traženi procjenitelj može zapisati i u obliku

$$a_0 + z\bar{X}\,,$$

gdje je naravno

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Dakle, jednadžbe (2.7) i (2.9) čine sustav od dvije linearne jednadžbe s dvije nepoznanice, a njegovim rješavanjem slijedi

$$a_0 = (1-z)E[m(\theta)] \quad \text{i} \quad z = \frac{n}{n + E[s^2(\theta)]/\mathrm{var}(m(\theta))}.$$

Na kraju ponovo dobijemo procjenu za $m(\theta)$ u obliku

$$z\bar{X} + (1-z)E[m(\theta)]\,,$$

slično kao i u (2.6). Primjetite da je faktor povjerenja z istog oblika kao i u N/N odn. P/Γ modelu.

Konačno, moramo uočiti da smo tri parametra u gornjem računu $E[m(\theta)], E[s^2(\theta)]$ te $\mathrm{var}(m(\theta))$ tretirali kao unaprijed poznata iako to u praksi nije slučaj. U Bayesovskim primjerima kao npr. N/N modelu ti parametri zaista jesu poznati jer nam je poznata apriori razdioba. Kako ovdje apriori razdiobu ne znamo, u praksi ove brojeve moramo procjeniti.

O procjeni $Em(\theta), Es^2(\theta), \mathrm{var}(m(\theta))$

Pretpostavit ćemo da imamo na raspolaganju podatke o više polica istog tipa. Drugim riječima, štete za i-tu policu iznose $X_{i,j}$, $j \geq 1$. Pri tom prepostavljamo da vrijede sljedeće tvrdnje

- i-tu policu opisuje par $(\theta_i, (X_{i,j}))$, gdje je sl. varijabla θ_i parametar rizika ili heterogenosti, a niz sl. varijabli $(X_{i,j})$ predstavlja niz šteta u danoj polici.,
- parovi $(\theta_i, (X_{i,j}))$, $i = 1, 2, \ldots$ su n.j.d. Posebno su θ_i, $i = 1, 2, \ldots$ n.j.d.
- razdioba od $X_{i,j}$ ovisi o slučajnom parametru θ_i,
- uz dano θ_i sl. varijable $X_{i,j}$ su n.j.d.

Katkad se govori da je i-ta polica jedna u **kolektivu rizika, eng. collective**.

Prepostavimo da nas zanima odrediti neto premiju za i-tu policu. Bayesovska statistika nas upućuje na korištenje uvjetnog očekivanja

$$m(\theta_i) = E(X_{i,1}|\theta_i)\,.$$

No kao što smo gore argumentirali, ovo je očekivanje općenito vrlo teško izračunati, ali se možemo poslužiti optimalnim linearnim procjeniteljem. On je ovisio o tri parametra: $E[m(\theta_i)], E[s^2(\theta_i)]$ i $\mathrm{var}(m(\theta_i))$. Uočite da ove veličine ne ovise o i. Nadalje veličine $m(\theta_i)$ i $s^2(\theta_i)$ možemo procijeniti iz uzorka uobičajenim procjeniteljima

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{i,j}$$

te

$$\hat{s}_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_{i,j} - \bar{X}_i)^2$$

Ako su nam dostupni podaci za m polica, $E[m(\theta_i)]$ i $E[s^2(\theta_i)]$ možemo procjeniti koristeći njihove prosječne vrijednosti

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N}\bar{X}_i = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n}\sum_{j=1}^{n}X_{i,j}$$

i

$$\bar{s}^2 = \frac{1}{N}\sum_{i=1}^{N}\hat{s}_i^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n-1}\sum_{j=1}^{n}(X_{i,j}-\bar{X}_i)^2.$$

Za $\mathrm{var}(m(\theta_i))$ razuman procjenitelj je

$$\frac{1}{N-1}\sum_{i=1}^{N}(\bar{X}_i - \bar{X})^2,$$

no može se pokazati da je on pristran stoga koristimo korigirani procjenitelj

$$\hat{v} = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{X}_i - \bar{X})^2 - \frac{1}{n}\bar{s}^2\,.$$

lako je ovaj procjenitelj nepristran ni on u praksi nije korišten jer ponekad može primiti negativne vrijednosti tako da varijancu $\mathrm{var}(m(\theta_i))$ zapravo procjenjujemo koristeći sljedeći pristrani procjenitelj

$$\max\{\hat{v}, 0\}.$$

Ako ponovo promotrimo vrijednosti faktora povjerenja

$$z = \frac{n}{n + E[s^2(\theta)]/\mathrm{var}(m(\theta))}$$

u ovom modelu, ponovo možemo uočiti da je z rastuća funkcija od n tj. duljine uzorka. S druge strane z pada kako raste izraz $E[s^2(\theta)]/\mathrm{var}(m(\theta))$, to je takodjer razumljivo, jer on mjeri odnos varijabilnosti procjenitelja za $m(\theta_i)$ tj. za konkretnu policu u odn. na varijabilnost procjenitelja za sve rizike u kolektivu. Iako smo naglasili da faktor povjerenja ne bi trebao ovisiti o prikupljenim podacima, u praksi on ovisi o njima jer vrijednost $E[s^2(\theta)]/\mathrm{var}(m(\theta))$ ne znamo, pa je procjenjujemo iz podataka na gore opisan način.

Na kraju uočite da je zbog korištenja kvadratne funkcije gubitka kod ocjene optimalnosti linearnog procjenitelja, jedino bilo bitno znati da $E[m(\theta_i)], E[s^2(\theta_i)]$ i $\mathrm{var}(m(\theta_i))$ ne ovise o indeksu i. Tako da smo isti račun mogli provesti i pod nešto oslabljenim prepostavkama na naš model

- i-tu policu opisuje par $(\theta_i, (X_{i,j}))$, gdje je sl. varijabla θ_i parametar rizika ili heterogenosti, a niz sl. varijabli $(X_{i,j})$ predstavlja niz šteta u danoj polici.,
- parovi $(\theta_i, (X_{i,j})),\ i = 1, 2, \ldots$ su medjusobno nezavisni. Parametri rizika $\theta_i,\ i = 1, 2, \ldots$ su n.j.d.
- razdioba od $X_{i,j}$ ovisi o slučajnom parametru θ_i, ali može ovisiti i o indeksu j
- uz dano θ_i sl. varijable $X_{i,j}$ su nezavisne, a $E(X_{i,j}|\theta_i)$ kao i $\mathrm{var}(X_{i,j}|\theta_i)$ ne ovise o j

Ovaj model se u literaturi ponekad naziva Bühlmannovim. On se od dosad korištenog modela razlikuje u sljedećem

- matrica šteta $((X_{i,j})_{j\geq 1})_{i\geq 1}$ se sastoji od nezavisnih redaka $(X_{i,j})_{j\geq 1})$ koji nisu nužno jednako distribuirani,
- u polici i smo jedino zahtjevali da su $E(X_{i,j}|\theta_i)$ kao i $\mathrm{var}(X_{i,j}|\theta_i)$ neovisne o j, no razdiobe od $X_{i,j},\ j \geq 1$, nisu nužno jednake.

Dakle i pod ovim slabijim pretpostavkama gornji rezultati ostaju neprom-jenjeni.

Empirijska Bayesovska teorija povjerenja uz varijabilni volumen rizika (model 2)

Prethodni odjeljak počeli smo s prepostavkom da su iznosi šteta u sukcesivnim godinama n.j.d. sl. varijable ako nam je poznat parametar rizika θ. No u praksi se obim posla mjeren npr. brojem prodanih polica ili naplaćenim premijama mijenja s vremenom.

Razumno je prepostaviti da nam je **volumen rizika** za godine $i = 1, \ldots, n, n+1$ poznat i iznosi P_i u i-toj godini. Nadalje poznate su nam i dalje ukupne štete u portfelju za prvih n godina, recimo u iznosima $Y_1, \ldots, Y_n$. Nas i dalje zanima na osnovu ovih podataka procjeniti očekivanu ukupnu štetu u sljedećoj godini $E(Y_{n+1}|\theta)$. Primjetite da nam je volumen rizika, ili intuitivno obim posla u nadolazećoj godini poznat i iznosi P_{n+1}.

Iz ovih podataka možemo dobiti novi niz sl. varijabli na sljedeći način

$$X_i = Y_i/P_i, \qquad i = 1, 2, \ldots$$

Dakle X_i predstavlja ukupne štete u i-toj godini, ali standardizirane, odn. podjeljene ukupnim volumenom rizika u toj godini. Naravno za očekivati je da ovakve X_i imaju sličnije razdiobe nego Y_i. Zato se naše nove prepostavke tiču upravo ovih varijabli i glase

- policu (ili dio portfelja) modelira par $(\theta, (X_i))$, gdje je parametar rizika θ nepoznat, a niz sl. varijabli (X_i) predstavlja niz šteta u danoj polici,
- razdioba od X_i ovisi o slučajnom parametru θ,
- uz dano θ sl. varijable X_i su nezavisne, ali ne nužno i jednako distribuirane.
- Veličine

  $$m(\theta) = E(X_i|\theta) \text{ i } s^2(\theta) = P_i \mathrm{var}(X_i|\theta)$$

  ne ovise o i.

Kada je obim posla konstantan kroz vrijeme, recimo $P_i = 1$ za sve godine, ovaj novi model se ne razlikuje od Bühlmannovog. Inače je u literaturi poznat kao Bühlmann–Straubov model.

Primjer Za ilustraciju gornjih pretpostavki, uzmimo npr. da P_i predstavlja broj prodanih polica u i-toj godini, te da su pojedinačne štete nezavisne kada nam je poznat njihov zajednički parametar rizika θ. Nadalje pretpostavite da očekivanje odn. varijanca pojedinačne štete iznose $m(\theta)$ odn. $s^2(\theta)$ Uočite da je u ovom primjeru

$$E(Y_i|\theta) = P_i m(\theta) \text{ i } \mathrm{var}(Y_i|\theta) = s^2(\theta) P_i$$

dok za normalizirane ukupne štete $X_i = Y_i/P_i$ sada vrijedi

$$E(X_i|\theta) = m(\theta) \text{ i } \mathrm{var}(X_i|\theta) = s^2(\theta)/P_i.$$

Posebno su dakle zadovoljene prepostavke gore uvedenog modela.

I ovdje je optimalan odgovor na problem procjene neto premije odn. očekivane vrijednosti od Y_{n+1} uz dane podatke, zapravo Bayesovski procjenitelj $E(Y_i|\theta) = P_i m(\theta)$. Kako nam je on nepoznat i tipično vrlo kompliciran, i ovdje ćemo potražiti optimalan linearan procjenitelj za $m(\theta)$ u srednjekvadratnom smislu. Drugim riječima tražimo konstante $a_0, a_1, \ldots, a_n$ tako da minimiziramo sljedeći izraz

$$E\left(m(\theta) - (a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n)\right)^2.$$

Možemo primjeniti metodu diferenciranja i traženja kritičnih točaka kao i prije da odredimo ove konstante.

Jednostavan, ali podulji račun pokazuje kako su optimalne konstante:

$$a_0 = \frac{E[m(\theta)]E[s^2(\theta)]/\mathrm{var}(m(\theta))}{\sum_{i=1}^{n} P_i + E[s^2(\theta)]/\mathrm{var}(m(\theta))},$$

$$a_k = \frac{P_k}{\sum_{i=1}^{n} P_i + E[s^2(\theta)]/\mathrm{var}(m(\theta))},\ k = 1, 2 \ldots .$$

Sad kad je poznat optimalan linearan procjenitelj za $m(\theta)$ oblika $a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$, provjerite da i njega možemo zapisati u obliku

$$z\bar{X} + (1 - z)E(m(\theta))$$

uz korištenje težinskog prosjeka

$$\bar{X} = \frac{\sum_{i=1}^{n} P_i X_i}{\sum_{i=1}^{n} P_i}$$

i faktor povjerenja

$$z = \frac{\sum_{i=1}^{n} P_i}{\sum_{i=1}^{n} P_i + E[s^2(\theta)]/\mathrm{var}(m(\theta))}.$$

Uočite da koeficijenti a_k, $k = 1, 2, \ldots$, nisu nužno jednaki. Izuzev u slučaju kad su volumeni rizika jednaki. Posebno, ako su svi P_i jednaki 1, dobijemo isto rješenje kao i prijašnjem modelu, no to ne iznenadjuje jer se tada i modeli potpuno podudaraju. Na kraju, primjetimo da rješenje opet ovisi o tri parametra $E[m(\theta)]$, $E[s^2(\theta)]$ i $\mathrm{var}(m(\theta))$. No, i njih moramo procijeniti, pri tom možemo koristiti iste ideje kao i u prethodnom modelu.

Prije procjene dakle nabrojimo osnovne podatke i uvjete pod kojima provodimo statističku analizu. Oni su u suštini isti kao i prije, jedino sad imamo više podataka

- i-ti rizik opisuje par $(\theta_i, (P_{i,j}), (Y_{i,j}))$, gdje je sl. varijabla θ_i parametar rizika ili heterogenosti, $P_{i,j}$ je volumen tog rizika u j-toj godini, a sl. varijabla $Y_{i,j}$ predstavlja štetu i-tog rizika u j-toj godini,
- za sl. varijable $X_{i,j} = Y_{i,j}/P_{i,j}$ vrijedi da su parovi $(\theta_i, (X_{i,j}))$, $i = 1, 2, \ldots$ medjusobno nezavisni. Parametri rizika θ_i, $i = 1, 2, \ldots$ su n.j.d.
- razdioba od $X_{i,j}$ ovisi o slučajnom parametru θ_i, ali može ovisiti i o indeksu j
- uz dano θ_i sl. varijable $X_{i,j}$ su nezavisne, ali $m(\theta_i) = E(X_{i,j}|\theta_i)$ kao i $s^2(\theta_i) = P_{i,j}\mathrm{var}(X_{i,j}|\theta_i)$ ne ovise o j

Veličine $m(\theta_i)$ i $s^2(\theta_i)$ možemo procijeniti iz uzorka sljedećim procjeniteljima

$$\bar{X}_i = \frac{\sum_{j=1}^{n} P_{i,j} X_{i,j}}{\sum_{j=1}^{n} P_{i,j}}$$

te

$$\hat{s}_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} P_{i,j} (X_{i,j} - \bar{X}_i)^2 \,.$$

Stoga procjenitelji za $E(m(\theta_i))$ odn. $E(s^2(\theta_i))$ iznose

$$\bar{X} = \frac{\sum_{j=1}^{N} \sum_{j=1}^{n} P_{i,j} X_{i,j}}{\sum_{j=1}^{N} \sum_{j=1}^{n} P_{i,j}}$$

odn.

$$\bar{s}^2 = \frac{1}{N} \sum_{j=1}^{n} \frac{1}{n-1} \sum_{j=1}^{n} P_{i,j} (X_{i,j} - \bar{X}_i)^2 .$$

Može se pokazati da su oni nepristrani.

Ostaje nam procijeniti $\mathrm{var}(m(\theta_i))$, no pokazuje se da je nepristran procjenitelj dan izrazom

$$\hat{v} = \frac{1}{P^*}\left[\frac{1}{Nn-1}\sum_{j=1}^{n}\sum_{j=1}^{n} P_{i,j}(X_{i,j}-\bar{X}_i)^2 - \frac{1}{N}\sum_{j=1}^{n}\frac{1}{n-1}\sum_{j=1}^{n} P_{i,j}(X_{i,j}-\bar{X}_i)^2\right],$$

gdje je

$$P^* = \frac{1}{Nn-1}\sum_{j=1}^{n}\sum_{j=1}^{n} P_{i,j}\left(1-\frac{\sum_{j=1}^{n} P_{i,j}}{\sum_{j=1}^{n}\sum_{j=1}^{n} P_{i,j}}\right).$$

I ovdje je moguće da $\hat{v}$ poprimi negativne vrijednosti stoga se koristimo razumnijim, ali pristranim procjeniteljem $\max\{\hat{v}, 0\}$.
Kako i naslućujemo, u slučaju da su svi $P_{i,j}$ jednaki 1, ove formule se potpuno podudaraju s onima koje smo izveli u Bühlmannovom modelu.

3. Jednostavan sustav iskustvenog utvrdjivanja premija

Sustav bonusa

Široko je rasprostranjen princip odredjivanja premija koji visinu premije dovodi u ovisnost o broju šteta ugovaratelja osiguranja u prethodnim godinama. On se posebno često primjenjuje kod osiguranja motornih vozila. Ovaj princip se naziva **sustavom bonusa** ili na engleskom **NCD - No Claim Discount**. U praksi dulji niz godina bez šteta, znači i veći popust za osiguranika kod ugovaranja premije. Jedna od poželjnih nuspojava ovog sustava za osiguravatelja je da obeshrabruje prijavljivanje malih šteta. Administriranje malih šteta, je s druge strane razmjerno veliki trošak za osiguravatelja. Uštede koje osiguravatelj ostvaruje na ovaj način omogućuju mu postavljanje nižih tj. kompetitivnijih premija.

Sustav bonusa čine dvije komponente:

- kategorije popusta, označene brojevima: $0, 1, 2, \ldots, k$, s pripadajućim popustom $0 \leq x_1 \leq x_2 \leq \cdots \leq x_k$.
- pravila prelaska iz jedne kategorije u drugu..

Kategorije tipično odgovaraju broju godina bez štete. Pravila mogu nakon prijave štete osiguranika prebaciti u kategoriju s nižim popustom, npr. za jednu ili dvije kategorije. U praksi popust osiguranika može ovisiti i o drugim poznatim faktorima rizika, npr. dobi, spolu i sl.

Primjer (Sustav bonusa s tri kategorije) Prepostavimo da imamo homogenu grupu osiguranika, te da su odredjene sljedeće kategorije u kojima postoci označavaju popust u odn. na punu premiju.

Kategorija	Popust
0	0 %
1	25%
2	40 %

Ako osiguranik prijavi štetu, spušta se za jednu kategoriju ako je to moguće ili ostaje u kategoriji 0 ako nije. I obrnuto, nakon svake godine bez prijave štete osiguranik se uspinje za jednu kategoriju, osim ako već nije u kategoriji 2, kada u njoj i ostaje.

U praksi se koriste sustavi s više kategorija nego u gornjem primjeru (npr. 5 ili 6), i kompliciranijim pravilima prelaska. Ono što nas zanima je tipično izračunati očekivanu štetu i vrijednost premija u homogenoj grupi osiguranika kroz dulji niz godina. Zato nam trebaju dodatne informacije. Prije svega potrebno je znati vjerojatnost da će osiguranici pretrpiti štetu u pojedinoj godini, a zatim i vjerojatnosti prelaska iz kategorije u kategoriju. One se obično zadaju u **matrici prijelaza** oblika

$$P = \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0k} \\ p_{10} & p_{11} & \cdots & p_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k0} & p_{k1} & \cdots & p_{kk} \end{bmatrix}.$$

Broj p_{ij} označava vjerojatnost prelaska iz stanja i u stanje j nakon jedne godine.

Čitaoc koji poznaje definiciju Markovljevih lanaca prepoznat će da je proces koji opisuje put osiguranika kroz kategorije upravo jedan takav lanac. U praksi je potrebno znati i inicijalnu razdiobu lanca, odn. u našem slučaju osiguranika po kategorijama, ona je odredjena vektorom vjerojatnosti $\pi^{(0)} = (\pi_0^{(0)}, \pi_1^{(0)}, \ldots, \pi_k^{(0)})$. Npr. ako su na početku svi osiguranici u kategoriji 0, imamo $\pi^{(0)} = (1, 0, \ldots, 0)$.

Ako označimo s $\pi^{(i)} = (\pi_0^{(i)}, \pi_1^{(i)}, \ldots, \pi_k^{(i)})$, razdiobu osiguranika po kategorijama nakon i godina. Lako se vidi da vrijedi

$$\pi_l^{(i)} = \sum_{j=0}^{k} \pi_j^{(i-1)} p_{jl},$$

ili u vektorskom zapisu

$$\pi^{(i)} = \pi^{(i-1)} P.$$

Iz teorije Markovljevih lanaca poznato je da pod relativno blagim uvjetima, razdioba $\pi^{(i)}$ teži k nekoj graničnoj razdiobi π, za $i \to \infty$. Iz $\pi^{(i)} = \pi^{(i-1)}P$, sada slijedi i da razdioba π zadovoljava

$$\pi = \pi P\,,$$

a za takvu razdiobu kažemo da je invarijantna za danu matricu prijelaza. U jeziku linearne algebre je vektor π svojstveni vektor za svojstvenu vrijednost 1 matrice P u odn. na množenje matricom zdesna.

Primjer (nastavak prethodnog) U bonus sustavu s tri kategorije pretpostavite da je vjerojatnost prijavljivanja štete 0.1, nadjite stacionarnu distribucija osiguranika po kategorijama.

$$P = \begin{bmatrix} 0.1 & 0.9 & 0 \\ 0.1 & 0 & 0.9 \\ 0 & 0.1 & 0.9 \end{bmatrix}.$$

Ako ispišemo sustav jednadžbi za $\pi = \pi P$ dobit ćemo

$$\begin{aligned} 0.1\pi_0 + 0.1\pi_1 &= \pi_0 \\ 0.9\pi_0 + 0.1\pi_2 &= \pi_1 \\ 0.9\pi_1 + 0.9\pi_2 &= \pi_2 \end{aligned}$$

Kako se radi o linearno zavisnim jednadžbama moramo iskoristiti i činjenicu da vektor π predstavlja razdiobu, tj, vrijedi

$$\pi_0 + \pi_1 + \pi_2 = 1\,.$$

Riješavanje sustava nam daje

$$\pi = \left(\frac{1}{91}, \frac{9}{91}, \frac{81}{91}\right).$$

Pokažite da granična razdioba, u slučaju kad je vjerojatnost prijavljivanja štete u nekoj godini 0.2, iznosi $\pi = (1/21, 4/21, 16/21)$.

Pogledajmo sada što se dogadja ako sustav bonusa primjenimo na heterogeni portfelj. Npr. portfelj osiguranja vozila u kojem osiguranike dijelimo na "dobre" i "loše" vozače. Podrobnija analiza većine korištenih sustava bonusa u praksi, pokazuje da u heterogenim portfeljima oni ne funkcioniraju kako su dizajnirani. Naime osiguranici ne plaćaju premije proporcionalno svojim očekivanima vrijednostima štete. Iako je matematički uvijek moguće postaviti broj kategorija i pravila tako da sustav zadovolji i zahtjev proporcionalnosti, u praksi se to ne čini jer bi takav sustav bio složen i teško razumljiv.

Primjer (nastavak prethodnog) Prepostavima da u sustavu s tri kategorije nositelje osiguranja možemo dijeliti na "dobre" i "loše vozače". Vjerojatnost prijavljivanja štete u jednoj godini je za dobre vozače je 0.1, dok za loše ona iznosi 0.2. Prepostavimo da je razdioba visine odštetnih zathjeva ista u obje skupine. Za očekivati je da premija loših vozača mora biti dvostruka u odn. na dobre. Prisjetite se da stacionarnu razdiobu dobrih vozača po kategorijama već znamo. Ako puna premija iznosi c, onda je očekivana vrijednosti prosječne premije plaćene u ovoj grupi osiguranika

$$\frac{1}{91}c + \frac{9}{91}0.75c + \frac{81}{91}0.6c = 0.619c.$$

Slično za loše vozače ona iznosi

$$\frac{1}{21}c + \frac{4}{21}0.75c + \frac{16}{21}0.6c = 0.648c,$$

dakle razlika je gotovo neznatna.

Razlozi za ovu pojavu su intuitivno razumljivi. Kao prvo, popusti su relativno mali, kao i broj kategorija, što ne omogućava preciznije razlikovanja dvije grupe vozača. Kao drugi razlog, uočite da je vjerojatnost šteta u obje grupe relativno mala, pa vozači i u jednoj i u drugoj imaju visoke vjerojatnosti uživanja maksimalne razine popusta.

Posljedice sustava bonusa na vjerojatnosti prijavljivanja šteta

Do sada smo prepostavljali da osiguranici prijavljuju sve nastale štete, no to ne mora biti tako. Ako se vratimo našem primjeru, i pretpostavimo punu premiju u iznosu 500, ona za druge dvije kategorije iznosi 375, odn. 300. U odluci da li će prijaviti štetu ili ne, osiguranik može voditi računa o budućim premijama.
Ako je npr. pripadao kategoriji 0, te ako ne bude imao više nezgoda u budućnosti, njegove premije će iznositi

$$500, 375, 300, 300, \ldots, \quad \text{ako prijavi štetu}$$

ili

$$375, 300, 300, 300, \ldots, \quad \text{ako ne prijavi štetu.}$$

Dakle, razlika budućih premija je 200, pa mu se ovako gledano ne isplati prijaviti štetu manju od tog iznosa.

Ako je vozač pripadao kategoriji 1, razlika premija će iznositi 275, a u kategoriji 2 ona će iznositi 75. Iz toga možemo zaključiti da će u slučaju nezgode, osiguranici u raznim kategorijama prijavljivati štete s različitim vjerojatnostima.

U gornjim izračunima, gledali smo razliku izmedju premija tokom sljedećeg perioda sve dok vozač ponovo ne dostigne kategoriju maksimalnog popusta. No osiguranik može promatrati i kraće periode, npr. ako očekuje dodatne štetu u bliskoj budućnosti. Period koji koristi osiguranik da bi razmotrio razliku premija izmedju prijavljivanja i neprijavljivanja štete zove se **osiguranikov horizont**.

Kako smo vidjeli, vjerojatnosti da osiguranik pretrpi gubitak nije jednaka vjerojatnosti prijavljivanja štete. O tome moramo voditi računa kad odredjujemo vjerojatnosti prelaska. Ako je poznata razdioba iznosa štete i horizont osiguranika za sve osiguranike u našem portfelju, za očekivati je da će racionalni osiguranici prijaviti štetu u iznosu X ako ona zadovoljava

$X > c_u$ = ušteda premija koju osiguranik unutar svog vremenskog horizonta postiže neprijavljivanjem štete.

U našem primjeru osiguranik iz kategorije 1, s vremenskim horizontom od bar dvije godine neće prijaviti štetu, ako ona iznosi manje od $c_u = 275$. Općenito je vjerojatnost da će osiguranik prijaviti štetu u danoj godini (i pri tom tipično promijeniti kategoriju na niže) jednaka

$$P(\text{ nezgoda })P(\text{ prijava } | \text{ nezgoda }),$$

gdje je

$$P(\text{ prijava } | \text{ nezgoda }) = P(X > c_u)\,.$$