

UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO

Finančna matematika – 1. stopnja

Neža Kržan

Statistika v kazenskem pravu

Delo diplomskega seminarja

Mentor:izr. prof. Jaka Smrekar

Ljubljana, 2022

KAZALO

1. Bayesova statistika	3
1.1. Bayesovo pravilo	3
1.2. Bayesovo posodabljanje	3
2. Primer - Taksi podjetja	4
3. Identifikacijski dokazi	5
3.1. Poenostavljena Bayesova analiza	5
3.2. Izpopolnjena Bayesova analiza	6
4. Drugi pristopi	7
4.1. Frekvence in edinstvenost	7
4.2. Naključno ujemanje in razmerja verjetnosti	8

1. BAYESOVA STATISTIKA

Bayesova statistika je statistična veja, ki nam s pomočjo matematičnih pristopov omogoča uporabo verjetnosti pri reševanju statističnih problemov. V svoje modele vključuje pogojno verjetnost, katero izračunamo z uporabo Bayesovega pravila.

Zlasti Bayesovo sklepanje razlaga verjetnost kot merilo verjetnosti ali zaupanja, ki ga lahko ima posameznik glede nastanka določenega dogodka. O nekem dogodku lahko že imamo predhodno prepričanje oziroma apriorno prepričanje, ki pa se lahko spremeni, ko se pojavijo novi dokazi. Bayesova statistika nam daje matematične modele za vključevanje naših apriornih prepričanj in dokazov za ustvarjenje novih prepričanj oziroma za pridobitev aposteriornega prepričanja, ki se lahko uporabi za kasnejše odločitve.

1.1. Bayesovo pravilo. Bayesovo sklepanje temelji na Bayesovim pravilom, ki izraža verjetnost nekega dogodka z verjetnostjo dveh dogodkov in obrnejnje pogojne verjetnosti. Pogojna verjetnost predstavlja verjetnost dogodka, glede na drug dogodek.

Definicija 1.1. Pogojna verjetnost dogodka H , glede na dogodek E , je

$$(1) \quad P(H|E) = \frac{P(H \cap E)}{P(E)},$$

ob predpostavki, da je $P(E) > 0$.

Formula (1) pove, da je verjetnost dogodka H ob pogoju, da se je zgodil dogodek E , enaka razmerju verjetnosti, da se zgodita oba dogodka in verjetnosti, da se je zgodil dogodek E .

Potem pogojno verjetnost uporabimo še v števcu formule (1) in dobimo Bayesovo pravilo:

$$(2) \quad P(H|E) = \frac{P(E|H) \times P(H)}{P(E)},$$

Verjetnost dogodka E lahko še razpišemo in dobimo:

$$(3) \quad P(H|E) = \frac{P(E|H) \times P(H)}{P(E|H)P(H) + P(E|\neg H)P(\neg H)}.$$

Obstaja še ena formulacija Bayesovega pravila, ki olajša izračune in je pogosto uporabljena pri Bayesovi analizi DNK dokazov:

$$(4) \quad \frac{P(H|E)}{P(\neg H|E)} = \frac{P(E|H)}{P(E|\neg H)} \times \frac{P(H)}{P(\neg H)}.$$

1.2. Bayesovo posodabljanje. Bayesovo pravilo se razlikuje od Bayesovega posodabljanja. Prvo je matematični izrek, drugo pa logična trditev, kako se sčasoma posodablja apriorne verjetnosti dokazov glede na novo zbrane dokaze oziroma prepričanja.

Bayesovo posodabljanje pravi:

Trditev 1.2. Če se dogodek E zgodi ob času $t_1 > t_0$, potem je $P_1(H) = P_0(H|E)$.

Ob času t_0 dogodku H dodelimo verjetnost $P_0(H)$; to se imenuje predhodna verjetnost oziroma apriorna verjetnost. Ko se zgodi dogodek E ob času t_1 , ki vpliva na naša prepričanja o dogodku H , Bayesovo posodabljanje pravi, da je potrebno apriorno verjetnost dogodka H v času t_1 enačiti z pogojno verjetnostjo dogodka H glede na dogodek E v času t_0 .

Recimo, da je dogodek H neka hipoteza oziroma prepričanje o zločinu in dogodek E dokazi, zbrani za ta zločin. Pri Bayesovem posodabljanju je videti, kot da je dokaz E nesporno resničen. Z drugimi besedami, predpostavka je, da moramo imeti po zbiranju dokazov E stopnjo zaupanja v E enako 1, torej če so dokazi zbrani v času t_1 , je $P_1(E) = 1$.

2. PRIMER - TAKSI PODJETJA

Za lažje razumevanje Bayesovega pravila si pogledjmo primer.

Obstajata dve taksi podjetji, Zeleni Taksi in Modri taksi, katerih vozila so pobarvana zeleno oziroma modro. Podjetje Zeleni taksi pokriva 85 odstotkov trga, podjetje Modri taksi pa preostanek. Predpostavimo še, da v okolici ni drugih taksij. V megletem dnevu taksi trči v mimoidočega pešca in ga poškoduje, vendar odpelje s kraja nesreče. Priča nesreče poroča, da je bilo vozilo modre barve. Priča ima prav le v 80 odstotkih primerov, kar pomeni, da je njegova zanesljivost enaka 0,8. Kolikšna je verjetnost, da je bil taksi, ki je povzročil nesrečo, modre barve glede na poročilo priče?

Vpeljimo oznake:

Z ... hipoteza, da je bil taksi zelen,

M ... hipoteza, da je bil taksi moder,

W_m ... dokaz, t.j. poročanje priče, da je bil taksi moder.

Problem je določiti pogojno verjetnost hipoteze M , ob pogoju, da je dokaz W_m resničen, torej $P(M|W_m)$.

Za uporabo Bayesovega pravila potrebujemo tri elemente: verjetnost dokazov glede na hipotezo, verjetnost dokazov in verjetnost hipoteze. Podjetje Zeleni taksi pokriva 85 odstotkov trga, zato je verjetnost $P(Z) = 0,85$. Po definiciji verjetnosti, je potem $P(M) = 1 - P(Z) = 0,15$. Vemo tudi, da ima priča v 80 odstotkih primerov prav, torej $P(W_m|M) = 0,8$ in $P(W_m|Z) = 0,2$. Izračunajmo še verjetnost dokaza $P(W_m)$:

$$P(W_m) = P(W_m|Z)P(Z) + P(W_m|M)P(M) = 0,2 \times 0,85 + 0,8 \times 0,15 = 0,29.$$

Sedaj lahko uporabimo Bayesovo pravilo:

$$P(M|W_m) = \frac{P(W_m|M)P(M)}{P(W_m|M)P(M) + P(W_m|Z)P(Z)} = \frac{0,8 \times 0,15}{0,29} = \frac{12}{29} \approx 0,41.$$

Verjetnost, da je bil taksi glede na pričanje dejansko modre barve, je precej majhna, tudi če ima priča v 80 odstotkih primerov prav. Razlog za to je, da je verjetnost M , ne glede na dokaze, majhna ($P(M) = 0,15$). Spodnja tabela kaže, da lahko s spreminjanjem verjetnosti M dobimo različne pogojne verjetnosti $P(M|W_m)$, pri čemer je zanesljivost priče nespremenjena:

$P(M)$	$P(Z)$	$P(W_m M)$	$P(M W_m)$
0,15	0,85	0,8	0,41
0,25	0,75	0,8	0,57
0,35	0,65	0,8	0,68
0,45	0,55	0,8	0,76
0,50	0,50	0,8	0,80
0,55	0,45	0,8	0,83
0,65	0,35	0,8	0,88
0,75	0,25	0,8	0,92
0,85	0,15	0,8	0,95

3. IDENTIFIKACIJSKI DOKAZI

V kazenskih zadevah želimo vedeti, ali je obtoženec kriv ali ne. Če imamo torej na voljo dokaze E , nas zanima pogojna verjetnost $P(kriv|E)$, pri čemer nam je lahko v pomoč Bayesovo pravilo. To v teoriji drži, čeprav je v praksi izračun verjetnostne krivde lahko preveč zapleten. Ampak z Bayesovim pravilom lahko ocenimo verjetnosti vmesnih trditev oziroma dokazov, ki so ključnega pomena za ugotavljanje obtoženčeve krivde. Najbolj pogosta uporaba Bayesovega pravila je pri ugotavljanju, ali je obtoženec vir sledi DNK-ja s kraja zločina.

3.1. Poenostavljena Bayesova analiza. Naj bo:

S ... trditev, da je obtoženec vir sledi DNK s kraja zločina;

M ... trditev, da se obtoženčev DNK ujema z DNK-jem s kraja zločina;

f ... funkcija pogostosti ujemanja DNK z DNK-jem s kraja zločina.

Želimo vedeti, kakšna je verjetnost S glede na M , tj. $P(S|M)$.

Bayesovo pravilo lahko uporabimo na naslednji način:

$$\frac{P(S|M)}{P(\neg S|M)} = \frac{P(M|S)}{P(M|\neg S)} \times \frac{P(S)}{P(\neg S)}.$$

Verjetnosti $P(S)$ in $P(\neg S)$ je težko oceniti, ker ne vemo kakšna je množica osumljencev. Smiselno bi bilo, da zanju upoštevamo interval predhodnih verjetnosti in ocenimo njihov vpliv na ne verjetnost trditve S in njene negacije. Nato moramo določiti vrednost $P(M|S)$, ki je običajno enaka ena - če bi obtoženec dejansko pustil sledove, bi laboratorijske analize pokazale ujemanje (to imenujemo lažno negativni rezultat); to je sicer poenostavitev, saj se lahko zgodi, da analize ne pokažejo ujemanja, čeprav je obtoženec pustil sledi. Potrebujemo še verjetnost $P(M|\neg S)$ (verjetnost, da se bo našlo ujemanje, če obtoženec ni vir sledi na kraju zločina). To je običajno enakovredno pogostosti ujemanja DNK-ja z DNK-jem s kraja zločina (tj.

f); tudi to je poenostavitev, saj se lahko zgodi, da obtoženec nima enakega DNK profila, vendar so laboratorijske analize pokazale, da ga ima (to imenujemo lažno pozitiven rezultat). Sledi:

$$\frac{P(S|M)}{P(\neg S|M)} = \frac{1}{f} \times \frac{P(S)}{P(\neg S)}.$$

3.2. Izpopolnjena Bayesova analiza. Poenostavljena Bayesova analiza ne upošteva možnosti lažno pozitivne laboratorijske analize. Za upoštevanje te možnosti, bomo namesto M uvedli spremenljivko M_p .

M_p ... poročano ujemanje laboratorijske analize;

M_t ... trditev, da obstaja dejansko ujemanje v DNK-ju;

$\neg M_t$... teditev, da obstaja neujemanje v DNK-ju.

Sledi:

$$P(M_p|\neg S) = P(M_p|M_t)P(M_t|\neg S) + P(M_p|\neg M_t)P(\neg M_t|\neg S).$$

Sedaj je $P(M_t|\neg S)$ enako f in zato $P(\neg M_t|\neg S)$ enako $1-f$. $P(M_p|\neg M_t)$ opisuje verjetnost lažno pozitivnih rezultatov laboratorija (oznaka FP) in $P(M_p|M_t)$ verjetnost resničnih pozitivnih rezultatov laboratorija (oznaka FN). Sledi:

$$P(M|\neg S) = [(1 - FN) \times f] + [FP \times (1 - f)].$$

Formula pokaže, da za parvilno oceno verjetnosti $P(M_p|\neg S)$ potrebujemo statistično oceno pogostosti profila DNK in stopnje napak laboratorijskih analiz, ki pa so redko na voljo.

Druga poenostavitev je, da predpostavimo, da je $P(M_p|S) = 1$, pri čemer ni upoštevana možnost lažnega negativnega rezultata. Kot zgoraj, imamo:

$$P(M_p|S) = P(M_p|M_t)P(M_t|S) + P(M_p|\neg M_t)P(\neg M_t|S).$$

Če je $P(M_p|S) = 1$, je $P(\neg M_p|S) = 0$ in $P(M_p|S) = 1 - FN$. Potem sledi, da je:

$$\frac{P(M_p|S)}{P(M_p|\neg S)} = \frac{1 - FN}{[(1 - FN) \times f] + [FN \times (1 - f)]}.$$

Za predstavo, kako stopnje napak vplivajo na razmerje verjetnosti, predpostavim, da je pogostost profila DNK 1 proti milijardi. Predpostavim tudi, da sta stopnji lažno pozitivnih in negativnih laboratorijskih rezultatov (tj. FP in FN) enako 0,01. Če je razmerje verjetnosti enako $\frac{1}{f}$, je milijarda. S formulo pa dobimo:

$$\frac{P(M_p|S)}{P(M_p|\neg S)} = \frac{1 - 0,01}{[(1 - 0,01) \times 0,000000001] + [0,01 \times (1 - 0,000000001)]} \approx 99.$$

Relativno majhne stopnje napak lahko bistveno zmanjšajo dokazno vrednost DNK dokazov, saj močno zmanjšajo razmerje verjetnosti; v našem primeru smo iz milijarde prišli na približno 100.

4. DRUGI PRISTOPI

Da bi ocenili moč Bayesove analize dokazov DNK, jo je koristno primerjati z nekaterimi drugimi pristopi.

4.1. Frekvence in edinstvenost. Predlagano je bilo s strani mnogih avtorjev, da je bolj naraven način za obravnavo verjetnosti uporaba naravnih frekvenc (angl. natural frequencies). Recimo, da je pogostost profila DNK f 1 proti 10 milijonov, predpostavimo tudi, da ima obtoženec enak DNK in da je začetna populacija osumljencev 100 milijonov ljudi. Zanima nas kakšna je verjetnost, da je obtoženec vir DNK-ja s kraja zločina.

Naj bo:

f ... pogostost profila DNK;

m ... velikost populacije osumljencev;

n ... število ljudi, ki imajo ustrezen DNK profil.

Po metodah, ki temeljijo na naravnih frekvencah, je potrebno izračunati, koliko ljudi z zadevnim profilom DNK je v populaciji osumljencev, tako da se f pomnoži z m . Če je posameznikov s takim profilom $n > 1$, je verjetnost, da je obtoženi vir, $\frac{1}{n}$. V primeru, ko je $n < 1$ in so frekvence še posebej majhne, je bolje raziskovati ali je profil DNK edinstven ali ne - ali poleg obtoženca obstajajo še drugi posamezniki z enakim profilom DNK. Temu pravimo metoda edinstvenosti (angl. uniqueness method).

S formulo binomske porazdelitve lahko izračunamo verjetnost, da se bo dogodek X , na primer profil DNK, pojavil k - krat v s - kratnem številu ponovitev, pri čemer ima dogodek X frekvenco $4f$. Želimo vedeti, kolikšna je verjetnost, da ima točno en posameznik ustrezen DNK profil, ob pogoju da ga ima vsaj en posameznik, torej:

$$P(n = 1 | n \geq 1) = \frac{P(n = 1 \cap n \geq 1)}{P(n \geq 1)} = \frac{m \times f \times (1 - f)^{m-1}}{1 - (1 - f)^m}.$$

m	f	$P(n = 1 n \geq 1)$	$P(S M)[\text{Bayes}]$
10 milijonov	1 v 100 milijonov	0,9	0,9
100 milijonov	1 v 100 milijonov	0,62	0,5
1 milijarda	1 v 100 milijonov	0	0,09
10 milijonov	1 v 1 milijardi	1	0,99
100 milijonov	1 v 1 milijardi	0,9	0,9
1 milijarda	1 v 1 milijardi	0,62	0,5
10 milijonov	1 v 10 milijardah	1	0,999
100 milijonov	1 v 10 milijardah	1	0,99
1 milijarda	1 v 10 milijardah	0,9	0,9

Tabela vsebuje primerjavo rezultatov, ki jih daje Bayesovo pravilo. Obe metodi se dokaj ujemata, vendar se tudi razlikujeta, saj v nasprotju z Bayesovim pravilom pri metodi edinstvenosti ni mogoče enostavno upoštevati številnih zapletov, kot je bil na primer vpliv stopnje laboratorijskih napak.

4.2. Naključno ujemanje in razmerja verjetnosti. Metoda verjetnost naključnega ujemanja(angl. random match probability) izraža možnost, da bi imel naključni posameznik, ki ni povezan z obdolžencem, ustrezeni DNK profil. Ta verjetnost je enaka pogostosti profila DNK. Težava tega pristopa je, da verjetnost naključnega ujemanja lahko predstavljena oziroma razumevana narobe.

Pogosto se to verjetnost interpretira na slednji način:

- (1) če je verjetnost naključnega ujemanja na primer 1 proti 100 milijonom, potem je verjetnost, da ima profil DNK drug posameznik in ne obdolženec 1 proti 100 milijonom;
- (2) ker je to zelo majhna verjetnost, mora biti tudi verjetnost, da je sled DNK pustil nekdo drug na kraju zločina in ne obdolženec, zelo majhna;
- (3) zato mora biti verjetnost, da je vir sledi DNK s kraja zločina obtoženec zelo velika, ampak znaša 1 proti 100 milijonov.

Takšno sklepanje je napačno in je znano kot tožilčeva zmota(angl. Prosecutor's fallacy). Sestavlja jo enačba

$$(5) \quad 1 - f = P(S|M).$$

Zmota se pojavi v koraku (2), ko je zamenjano $P(M|\neg S)$ s $P(\neg S|M)$ in predpostavljeno, da sta obe verjetnosti enaki f .