„It all depends on your prior!”

*Bayes-statisztika és hasonlóságelemzés*

Thomas Bayes a XVIII. században a valószínűségszámítás fontos tételét fogalmazta meg, melyet azonban életében nem publikált. Talán azért, mert sejthette, nem kis felzúdulást kelt majd a valószínűség fogalmának, illetve az ok-okozati kapcsolatok vizsgálatának újszerű megközelítési lehetőségeivel. Bayes személyes motivációja persze pillanatnyilag az anekdota homályában marad, ugyanakkor az tény, hogy a ma már Bayes-tételként ismert összefüggés publikálását fontos viták követték – s tény az is, hogy a tétel maga ma már lényegében középiskolai tananyag.

A feltételes valószínűség fogalmának és a teljes valószínűség tételének birtokában formálisan egyszerűen adódik a Bayes-tétel, mely szerint
Amennyiben azonban a feltételes valószínűség fogalmának ok-okozati jelentést tulajdonítunk, úgy az összefüggés lényegében a következmény vizsgálatából enged az ok valószínűségére következtetnünk. Például[[1]](#footnote-1) így lehet egy egészségügyi gyorsteszt eredményéből arra vonatkozó következtetést levonni, hogy a páciens milyen eséllyel fertőzött valamely adott vírussal.

Ez a megközelítés magában hordozza már ugyanakkor azt a fogalmi ellentmondást, ami végső soron a frekventista és bayesiánus statisztikai módszerek mögött ma is húzódik, s ami történetesen a valószínűség fogalmának legalapvetőbb értelmezése. Klasszikus értelemben valószínűségi kérdéseket vethet fel egy kocka vagy érme feldobása, hiszen ezek azonos körülmények között megismételhető események, s a valószínűség fogalma az egyes bekövetkező események relatív gyakoriságának határértéke. Az viszont, hogy holnap esni fog-e az eső, vagy melyik jelölt nyer egy szavazáson, nem ilyen jellegű kérdések – mégis, valahogy természetesnek adódik, ha ezek esetében esélyről, valószínűségről hallunk.

Az időjárás, illetve a választások kimenetele elvileg lehetne egy teljesen determinisztikus rendszer, ahol a jelenlegi nem-tudásunk az egyetlen tényező, ami meggátol bennünket abban, hogy a valódi eredményt előrejelezzük. Vagyis nincs itt semmi valószínűségi kérdés – pusztán nem tudunk eleget. Márpedig, ha ez a helyzet, akkor megkérdőjelezhető, szabad-e itt a valószínűség objektív fogalmát alkalmazni, hiszen igen könnyen lehet még tovább fokozni a szubjektív tudás figyelembe vételéből származó furcsaságokat.

Vegyünk ehhez egy pénzérmét. Legyen ez szabályos, és pillanatnyilag zárjuk ki, hogy feldobás után esetleg az élén megállhatna. Mielőtt feldobjuk, s ebben mindenki egyet mer talán érteni, mind a fej, mind az írás valószínűsége 0,5. Most én feldobom ezt az érmét, és mielőtt bárki láthatná, melyik oldalára is esett, gyorsan letakarom a tenyeremmel. Hogyan alakulnak ezek után a valószínűségek?

Szubjektíve minden továbbra is 0,5-nek tűnik, nem? Hiszen vagy fej, vagy írás; és ez most még csak nem is a Mátyás királyról[[2]](#footnote-2) szóló vicc. Más nézőpontból azonban ez már nem valószínűség kérdése, hiszen az esemény megtörtént, csak legfeljebb mi nem tudunk eleget a körülményekről[[3]](#footnote-3), hogy a pontos kimenetelt megmondhassuk. De ha valaki készített volna egy nagysebességű videofelvételt az érme feldobásáról, a lassított visszajátszás alapján láthatná, hogyan pörög az érme a levegőben, s akár még azt a rövidke pillanatot is elkaphatta volna, mielőtt a már asztalra érkezett érmét a kezemmel le tudom takarni. Vagy, ha paraván mögött, géppel történik is a feldobás; az érme kezdeti helyzete, a ráható erők pontos nagysága és iránya, illetve a gravitáció, a légellenállás és minden egyéb releváns paraméter figyelembe vételével akár ki is számolhatná, mi kell, hogy a kimenetel legyen.

Vagyis innen nézve ez egy biztos esemény; s ugyanez lenne a helyzet akkor is, ha az általam feldobott, majd tenyérrel letakart érme állását én meglesném, de az olvasónak nem árulnám el. Pontosan tudnám a kimenetelt, míg mindenki más találgathatna, és száz véletlenszerűen tippelő embernek nagyjából a fele járna sikerrel.

A statisztika nagyon szorosan épül a valószínűségszámításra, s alapvető (tudományos) elvárás, melyet a frekventista megközelítés igyekszik is kielégíteni, hogy a következtetések, amelyeket a statisztika alkalmazásával levonhatunk, objektívek legyenek. Jogosan adódik tehát az igény, hogy a fentebb bemutatott és ahhoz hasonló bizonytalanságokat ki kellene zárni a valószínűség fogalmából, hogy ilyesfajta szubjektivitás ne rondíthasson bele a tudomány objektív működésébe.

Érdemes ugyanakkor Bayes tételének állítását egy, az előzőnél konkrétabb megfogalmazásban is megvizsgálni (*H0* és *H1* alternatív hipotézisek helyességét jelölik, *K* egy konkrét, elvégzett kísérlet eredménye):
Ekkor ugyanis módszert kapunk arra, hogy közvetlenül meghatározzuk, milyen mértékben támasztja alá valamely kísérlet a meglévő hipotézisünket – s ez az, ami miatt a tudomány alapvetően alkalmazza a statisztikát. Ennek a kérdésnek a közvetlen megválaszolásához ugyanakkor fel kell használni azt, hogy – a kísérlet elvégzése előtt, s annak tapasztalatai nélkül – milyen esélyt tulajdonítunk a hipotézis helyességének.

*P(H*0*)* és *P(H*1*)* tehát azok a kritikus valószínűségek, amelyek értelmezésénél a frekventista és a bayesiánus statisztika összetűzésbe kerül. Ezek a kiindulási – *a priori* – valószínűségek viszik a megkérdőjelezhető szubjektivitást a következtetési folyamatba, s képesek a következtetéseket esetleg helytelen irányba is befolyásolni. Ugyanakkor az alkalmazásukkal olyan szakértői információk aknázhatók ki, amelyektől a frekventista megközelítés eltekint, s a nehezen megismételhető, valós helyzetekben pragmatikus döntési mechanizmust kínál a bayesiánus módszer.

Érdemes figyelembe venni azt is, hogy a bayesiánus módszer által eredményül szolgáltatott *P(H*0|*K)* értékét a frekventista statisztika bizonyos értelemben kerülő úton közelíti meg, éppen annak kihasználásával, hogy *P(K*|*H*i*)* értékeinek számítása a különböző eloszlások esetében könnyű feladat. Az alábbi ábra a konfidencia-intervallum fogalmát szemlélteti.



1. ábra: Binomiális eloszlás értéktáblázata (*n*=100),
oszloponként a legszűkebb 99% (kék), 90% (piros) és 75% (sárga) tartomány színezésével.
*Forrás: MS Excel, Pitlik M.*

Ha például az a kérdés, az ország lakosságának (kereskedelmi célcsoportban) mekkora része kedvel egy bizonyos csokoládét, akkor a közvélemény-kutató által telefonon vagy személyesen felkeresett válaszadók által alkotott mintában mért relatív gyakoriságból (*k/n*, vízszintes vonal) kell következtetést levonni az országos eredményekre. Ez esetben – egy, a fentihez hasonló – ábrán a vízszintes tengelyen kékkel jelölt és vastagított tartomány az, ahol azt mondhatjuk, hogy *amennyiben* az országban ennyi a csokoládét kedvelők aránya, *akkor* legalább 99% eséllyel elhihető, hogy valóban ebből a sokaságból kaptuk a *k/n* értéket produkáló mintánkat. Ugyanakkor a színes tartományok megállapítása függőlegesen történik, vö. *P(K*|*H*i*).*

Amennyiben valóban arra a kérdésre kíván az ember választ kapni, a kísérlet következtében hogyan értékelje a hipotéziseinek valószínűségét, nagy szükség van „jó” *priori*-eloszlásokra[[4]](#footnote-4). A hipotézisek előzetes valószínűségének meghatározása történhet korábbi mérések, adatok alapján, illetve az elemzést végző szakértő további szubjektív mérlegeléseit is alapul véve. Mindaddig, amíg a felhasznált információk mindegyike alaposan dokumentált, utólag is ellenőrizhető, szükség esetén korrigálható a bayesiánus módszerrel végzett hipotézisvizsgálat.

Ebből a szempontból kifejezetten jól illeszkedhet a Bayes-statisztika gondolatvilágába a hasonlóságelemzés (COCO) alkalmazása is. A hasonlóságelemzés módszertana ugyanis – hasonlóan a bayesiánus statisztikához – masszív számítástechnikai apparátusra épít, és az emberi szubjektivitásból származó, de mindenkor ellenőrizhető és szükség esetén felülbírálható keretfeltételek (irányvektorok) között mesterségesintelligencia-alapú becsléseket alkalmaz.

A fentebb megkezdett példa (csokoládé népszerűsége) alapján a COCO a következőképpen lehet képes releváns *a priori* információk előállítására. Legyenek ismertek a közvélemény-kutatás végrehajtása előtt is már bizonyos töredékes, „*nem-feltétlenül-reprezentatív*” adatok, melyek a csokoládé kedveltségére, keresletére vonatkoznak. Lehetnek ezek minimalista felmérések (internetes kérdőív), bizonyos boltok vásárlói adatai etc. A hasonlóságelemzéshez ezek alapján egy objektum-attribútum mátrix (OAM) felállítása szükséges.

Az OAM sorfejlécén foglalnak helyet az egyes töredékes minták (objektumok), az oszlopfejlécen pedig azok a jellemzők (attribútumok, *X*i), amelyek leírják a minta nem-reprezentativitását, vagyis például a minta és az országos átlag közötti különbséget életkor, nemi eloszlás, kereset, lakóhely típusa etc. tekintetében. Az utolsó oszlop (*Y*) az adott mintában a csokoládé relatív kedveltségére vonatkozó adat.

A fenti OAM méretfüggetlenítése szükséges a hasonlóságelemzés végrehajtása előtt. Ez azt jelenti, hogy minden egyes attribútumon belül az értékeket sorszámozni kell, attól függően csökkenő vagy növekvő sorrendben, hogy milyen szakértői információ / szubjektív tudás van az elemző birtokában arra nézve, hogy az adott attribútum minél nagyobb értéke a csokoládé kedveltségének minél nagyobb vagy kisebb értékével jár együtt. (Például, ha a csokoládéról ismert, hogy idősebbek körében kedveltebb, akkor a minta életkorának országos átlagtól való felfelé eltérését (attribútum) csökkenő sorrendben kell figyelembe venni.)

A hasonlóságelemzés során minden attribútum esetében meghatározásra kerül egy monoton csökkenő lépcsősfüggvény, melynek egyes értékei azt fejezik ki, az adott attribútum adott szintje milyen mértékben magyarázza a csokoládé kedveltségének felmért adatát az összehasonlítandó minták között. A hasonlóságelemzési célfüggvény az eredetileg felmért *Y*-adatok és lépcsősfüggvények alapján előállított objektumonkénti additív becslések eltérésének (delta) minimalizálása.

Végeredményben, alacsony hiba esetén, kijelenthető, hogy a különböző töredékes mintákon mért kedveltségi adatok közötti különbségek jól leírhatók a minták demográfiai eltérései alapján, így az országos kedveltségre vonatkozó becslés nyerhető minden attribútum esetén a megfelelő szélső lépcsőértékek összegzésével. Ugyanis az egyes attribútumok rendre az adott minta és az országos átlag eltéréseinek sorszámadatait tartalmazták, így ezen lépcsőértékek egy olyan ideális minta kedveltségi adatára (*Y*) adnak becslést, amelynek az országos demográfiai paraméterektől minden attribútum esetében minimális az eltérése.

Végeredményben a termék országos kedveltségére vonatkozó pontbecslés állítható elő, amelyet paraméterként lehet már használni egy megfelelő várható értékű béta-eloszlás megkonstruálására, s így a konkrétan megrendelt országos reprezentatív közvélemény-kutatás adataival való összevetésre. Az így felállított eljárás minden lényeges ponton támaszkodik a rendelkezésre álló szakértői adatokra (feltételezve azok viszonylagos pontosságát, hiszen az elemzőnek elemi érdeke, hogy ne használjon olyan *a priori* információkat, amelyek helytelenek). Ezzel párhuzamosan a relatíve számításigényes mesterséges intelligencia és bayesiánus megközelítések alkalmazása mellett minden felhasznált eleme az elemzésnek jól dokumentált és nyomon követhető, vagyis az elemzéstől elvárható objektivitást alapvetően támogatja.

**Források:**

Antal, É. (2010). *Bayes típusú problémák.* (szakdolgozat) ELTE TTK

Hunyadi, L. (2011). Bayesi gondolkodás a statisztikában. *Statisztikai Szemle, 89*(10-11). pp. 1150-1171.

Pitlik, L. (2014). *My-X Team, an Innovative „Idea-Breeding-Farm”*. Innoreg KMRIÜ

Pitlik, M. (2017). MS Excel makróval támogatott valószínűségi adatok (vizualizáció)

Eisner, J. és Johnson, P. írásai az alábbi internetes oldalon:
<https://www.quora.com/For-a-non-expert-what-is-the-difference-between-Bayesian-and-frequentist-approaches>

1. A híres Monty Hall-paradoxon feloldása is lehetséges a Bayes-tétel alkalmazásával. Ez egyúttal rávilágíthat arra is, hogy a tételben foglalt lehetőség, miszerint a következményekből az okokra visszakövetkeztethetünk, nem feltétlenül triviális, és nem is feltétlenül tűnik mindig természetesnek. [↑](#footnote-ref-1)
2. „– Mekkora az esélye annak, hogy Mátyás király szembejön velünk a folyosón?
– Ötven százalék, mert vagy igen, vagy nem!” [↑](#footnote-ref-2)
3. A dolog szegről-végről még Schrödinger macskájának is rokona. [↑](#footnote-ref-3)
4. A Bayes-tétel megfogalmazható folytonos valószínűségi változó esetén is, és például számolási szempontból kifejezetten alkalmas az ún. béta-eloszlások alkalmazása, mert béta *priori*-eloszlás alkalmazása esetén a *posteriori*-eloszlás is ilyen típusú lesz. [↑](#footnote-ref-4)