EÖTVÖS LORÁND TUDOMÁNYEGYETEM
TERMÉSZETTUDOMÁNYI KAR

A Bayes statisztika alkalmazása mindennapi helyzetekben

Készítette:
Dobreán István
Matematika BSc, elemző szakirány

Témavezető:
Dr. Vancsó Ödön
Matematikatanítási és Módszertani Központ

Budapest, 2020

NYILATKOZAT

Név: Dobreán István

ELTE Természettudományi Kar, szak: Matematika BSc

NEPTUN azonosító: UTBLHA

Szakdolgozat címe:
A Bayes statisztika alkalmazása mindennapi helyzetekben

A **szakdolgozat** szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló szellemi alkotásom, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 2020.12.29.

Dobreán István

a hallgató aláírása

Köszönet

Szeretnék köszönetet mondani a témavezetőmnek, Dr. Vancsó Ödönnek, a rengeteg segítségért és iránymutatásért, amivel a dolgozatom elkészültét segítette, még ebben a rendkívüli, nehéz időszakban is. Köszönöm továbbá családtagjaimnak és szeretteimnek is a támogatást és a türelmet az egész egyetemi tanulmányaim, és a szakdolgozat megírása alatt is.

Bevezetés

A szakdolgozatom témája a Bayes-statisztika bemutatása mindennapi alkalmazásokon keresztül. A statisztikában két fő irányzat létezik, a klasszikus, illetve a Bayes-statisztika. Mindkét módszer már évtizedek óta létezik, azonban az oktatásban, azon belül is főleg a felsőoktatásban a klasszikus statisztika van előtérben. Mindkét módszer jelentős képviselő táborral rendelkezik.

1997-ben egy vita is keletkezett arról, hogy a felsőoktatásban mekkora szerepet kellene adni a Bayes, illetve a klasszikus statisztikának. A két módszer képviselői leírták érveiket módszereik mellett. A klasszikus statisztika fő képviselője a vitában D.Moore professzor volt. Szerinte csak a klasszikus statisztikát kellene oktatni a felsőoktatásban.

Érvei között szerepelt, hogy a bayesi módszereket a gyakorlatban keveset használják, nincsenek kialakítva sztenderd módszerek és a feltételes valószínűség fogalmának nehézsége.

Ezzel szemben a másik oldal a Bayes statisztika felsőoktatásban való oktatását képviselte. Szerintük a fogalmi tisztázás miatt lenne fontos a Bayes statisztika oktatása. Továbbá azt állítják, hogy szerintük a klasszikus statisztika megzavarja a diákokat, és nehéz megérteni, mivel inverz gondolkodás szükséges hozzá.

A vita óta egy harmadik álláspont is felmerült, miszerint az imént említett két álláspontot párhuzamosan kellene oktatni. Ezáltal a diákok mindkét módszert láthatnák egymás mellett (Vancsó, 2005).

Jelen dolgozatomban a Bayes-statisztikát szeretném bemutatni alkalmazásokon keresztül.

Tartalomjegyzék

1. Alapfogalmak **6**
1.1. Bevezetés . . . 6
1.2. A feltételes valószínűség és a Bayes-tétel . . . 6
1.2.1. Feltételes valószínűség . . . 6
1.2.2. Bayes-tétel . . . 6
1.3. Példafeladat . . . 7
1.3.1. A megadott adatok . . . 7
1.3.2. Megoldás . . . 7

2. A koronavírus és a tesztek hatékonysága **9**
2.1. Bevezető . . . 9
2.1.1. Specificitás . . . 9
2.1.2. Érzékenység . . . 9
2.1.3. Fontos megjegyzés . . . 9
2.2. A tesztek . . . 10
2.2.1. A PCR teszt . . . 10
2.2.2. Rapid antitest teszt („gyorsteszt") . . . 10
2.2.3. Rapid antigén teszt . . . 10
2.2.4. Megjegyzés a tesztek paramétereiről . . . 10
2.3. A tesztek hatékonysága . . . 11
2.3.1. A PCR teszt . . . 11
2.3.2. Rapid antitest teszt („gyorsteszt") . . . 12
2.3.3. Rapid antigén teszt . . . 12
2.4. A tesztek megbízhatósága . . . 13
2.4.1. A PCR teszt . . . 13
2.4.2. Rapid antitest teszt („gyorsteszt") . . . 14
2.4.3. Rapid antigén teszt . . . 14
2.5. Több teszt elvégzése . . . 14
2.5.1. Gyorsteszt után PCR teszt . . . 14
2.5.2. Gyorsteszt után rapid antigén teszt . . . 15
2.5.3. PCR teszt után rapid antigén teszt . . . 15
2.6. Esetszám változtatás . . . 15
2.6.1. Ha kevesebb a ténylegesen fertőzött . . . 16
2.6.2. Ha több a ténylegesen fertőzött . . . 16
2.7. Kérdőív a koronavírus tesztek megbízhatóságáról . . . 16
2.7.1. Kérdések . . . 16
2.7.2. A kitöltők adatai . . . 17
2.7.3. Az eredmények . . . 17

2.7.4. Az eredmények kiértékelése 18
2.7.5. A többi teszt 18
2.8. Konklúzió 19

3. A hanglemez kereskedő **20**
3.1. Alapprobléma 20
3.2. A Bayes-elvről általánosan 20
3.2.1. A Bayes elv diszkrét esetben 20
3.2.2. A Bayes tétel folytonos esetben 21
3.2.3. A Bayes elv folytonos esetben 21
3.3. A feladat megoldása 21
3.4. n változtatása 22
3.4.1. Ha ingyenes a mintavétel 23
3.4.2. Ha a mintavétel nem ingyenes 23

4. A gyümölcskereskedő példája **26**
4.1. Alapprobléma 26
4.2. Feladatmegoldás 26
4.2.1. Becslés a keresletre 26
4.2.2. Veszteségfüggvény 27
4.3. Mennyi is az annyi? 28
4.3.1. A nyereség kiszámítása 28
4.3.2. Átlagkereslet szerinti nyereség 29

5. A gombák **30**
5.1. Alapprobléma 30
5.2. Naive Bayes osztályzás 30
5.2.1. Függetlenség 30
5.2.2. Az osztályozásról 30
5.3. A feladat 31
5.3.1. Az adathalmaz 31
5.3.2. A módszer bemutatása 32
5.3.3. A módszer végrehajtása a teljes adathalmazon pythonban . . . 33

6. Zárszó **34**
6.1. Végül egy izgalmas kérdés 34
6.2. A feladat egy lehetséges megoldása 34
6.3. Megjegyzés a használt értékekhez 36
6.4. Végszó 36

1. fejezet

Alapfogalmak

1.1. Bevezetés

A Bayes statisztika és a klasszikus statisztika közötti különbségek már évtizedek óta jelen vannak, nagy történelmi hátérrel rendelkeznek. A két irányzat alapvetően egymással szemben áll. A szembenállás lényegében filozófiai alapú. A Bayes-statisztikában fontos beszélnünk két alapfogalomról, melyek a következőek:

- Prior eloszlás: A kísérlet elvégzése előtti az illető szubjektív megítélése és informáltságán alapuló becslés.
- Poszteriori eloszlás: A tapasztalati statisztika alapján a priori eloszlás módosulásával létrejött új eloszlás.

1.2. A feltételes valószínűség és a Bayes-tétel

A következő részben egy rövid elméleti összefoglaló következik a továbbiakban használni kívánt jelölésekről, definíciókról és tételekről, Wickmann: Bayes Statisztika (1999) című könyve alapján.

1.2.1. Feltételes valószínűség

E_2 esemény E_1 feltétel melletti feltételes valószínűsége a következő:

$$P(E_2|E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)}$$

$P(E_1) > 0$ esetén (Wickmann 1999).

1.2.2. Bayes-tétel

Ha E_i egy teljes eseményrendszer, akkor a feltételes valószínűség értelmében (Wickmann 1999):

$$P(E_i|B) = \frac{P(B|E_i)P(E_i)}{P(B)} = \frac{P(B|E_i)P(E_i)}{\sum_{j=1}^{J} P(B|E_j)P(E_j)}$$

1.3. Példafeladat

Nézzünk meg egy példát a Bayes-tétel és a feltételes valószínűség alkalmazására. Adott egy teszt, ami meg tudja állapítani 0.8 pontossággal egy sofőrről, hogy a vérében az alkohol szintje milyen magas. Tehát a teszt pozitívat jelez az esetek 80%-ában, ha a sofőr ittas, és negatívat szintén az esetek 80%-ában, ha nem ittas. Ha pozitív lesz az első teszt, akkor egy másodikat is alkalmaznak. A második tesztről pedig azt tudjuk, hogy ha a sofőr nem ittas, akkor biztosan negatívat mutat, ha pedig ittas, akkor csak az esetek 10%-ában hibázik. Tegyük fel, hogy a rendőrök által megállított sofőrök negyede ittas. Határozzuk meg annak a valószínűségét, hogy valaki ténylegesen ittas, feltéve, hogy mindkét tesztje pozitív. (Migon és Gamerman, 1999)

1.3.1. A megadott adatok

Jelöljük θ-val a sofőr állapotát. $\theta = 0$, ha a sofőr nem ittas és $\theta = 1$, ha ittas. Legyen $x_i = 0$, ha az i-edik teszt negatív, és $x_i = 1$, ha az i-edik teszt pozitív. ($i = \{1, 2\}$) A feladat alapján a következőket ismerjük:

Mivel az első teszt pontossága 0.8:

- $P(x_1 = 1|\theta = 1) = 0.8$
- $P(x_1 = 0|\theta = 0) = 0.8$

Mivel a második teszt mindig kiszűri a nem ittasakat:

- $P(x_2 = 0|\theta = 0) = 1$
- $P(x_2 = 1|\theta = 0) = 0$

Mivel a második teszt csak az esetek 10%-ában hibázik az ittasoknál:

- $P(x_2 = 1|\theta = 1) = 0.9$
- $P(x_2 = 0|\theta = 1) = 0.1$

A kérdés: $P(\theta = 1|x_1 = 1, x_2 = 1) =?$

1.3.2. Megoldás

Először számoljuk ki annak a valószínűségét, hogy valaki ittas, feltéve, hogy az első teszt pozitív.

$$P(\theta = 1|x_1 = 1) = \frac{P(\theta = 1 \cap x_1 = 1)}{P(x_1 = 1)}$$

Először számoljuk ki $P(x_1 = 1)$-et. Tudjuk a feladatból, hogy a rendőrök által megállított sofőrök negyede ittas. Így:

$$P(x_1 = 1) = 0.8 \cdot 0.25 + 0.2 \cdot 0.75 = 0.35$$

Így a Bayes tétel segítségével:

$$P(\theta = 1|x_1 = 1) = \frac{0.8 \cdot 0.25}{0.8 \cdot 0.25 + 0.2 \cdot 0.75} = 0.571$$

Ebből az következik, hogy, ha a teszt pozitív, akkor annak az esélye, hogy valaki nem ittas: $P(\theta = 0|x_1 = 1) = 1 - 0.571 = 0.429$. Az eddig kiszámoltakból kiindulva nézzük meg, hogy mekkora az esélye, hogy valaki tényleg ittas, ha a második tesztje is pozitív.

$$P(\theta = 1|x_1 = 1, x_2 = 1) = \frac{P(x_1 = 1, x_2 = 1|\theta = 1) \cdot P(\theta = 1)}{P(x_1 = 1, x_2 = 1)}$$

A számlálóban azok az esetek találhatóak, amikor mindkét teszt pozitív és valóban ittas a sofőr, a nevezőben pedig általánosan mekkora az esélye, hogy mindkettő teszt pozitív. Számoljuk ki a nevezőt:

$$P(x_1 = 1, x_2 = 1) = 0 \cdot (0.429) + 0.9 \cdot (0.571) = 0.514$$

Így a Bayes-tétel alapján:

$$P(\theta = 1|x_1 = 1, x_2 = 1) = \frac{0.9 \cdot 0.571}{0.514} = 1$$

Ezek alapján láthatjuk, hogy 100% az esélye, hogy valaki ténylegesen ittas, ha mindkét teszt pozitívot mutat. Ez az eredmény szépen kijött a Bayes-tétel alkalmazásával, azonban érdemes megjegyezni, hogy számolás nélkül is rá lehet jönni, a józan ész alapján a megoldásra. Ugyanis ha a második teszt pozitív, akkor biztosan ittas a sofőr, így számolásra nincs szükség. Ezt onnan tudjuk, hogy a feladat azt írja, hogy a második teszt a nem ittasakra biztosan negatív eredményt hoz. Ez eléggé idealizált helyzet, ugyanis a valóságban nem lehet 100% pontossággal kiszűrni az ittasakat.

2. fejezet

A koronavírus és a tesztek hatékonysága

2.1. Bevezető

A 2020-as év egyik legmeghatározóbb témája a koronavírus volt. Ez a téma mindenkit érint az ország- és a világ nagy részén is. Mivel jelenleg is zajlik a járvány, így teljesen pontos adatokkal nem tudunk még számolni, csak közelítő értékekkel. Nézzük meg, a Bayes-statisztika milyen módon hasznosítható ezzel a témával kapcsolatban.

A legtöbb ember kíváncsi arra, hogy mekkora a valószínűsége annak, hogy koronavírusos, feltéve, hogy már tesztelték is őt, és a teszt pozitív lett. Több fajta tesztelési módszer is létezik, mindegyik más-más előnnyel és hátránnyal rendelkezik. A továbbiakban három különböző tesztről fogunk beszélni. A valóságban több fajta teszt is létezik, viszont az alábbi három a leggyakoribb, így ezekkel fogunk foglalkozni. Azonban mielőtt bemutatnánk a teszteket, két alapfogalmat tisztázni kell.

2.1.1. Specificitás

„Annak a valószínűsége, hogy a diagnosztikus teszt értéke negatív lesz egy olyan páciensen, akiben nem áll fenn a vizsgált betegség. A specificitás tehát azt jellemzi, hogy a teszt milyen megbízhatóan azonosítja azokat, akikben nem kóros a vizsgált paraméter."(Fidy & Makara, 2005)

2.1.2. Érzékenység

„Annak a valószínűsége, hogy a diagnosztikus teszt értéke pozitív lesz egy olyan páciensen, akiben fennáll a betegség. A szenzitivitás azt jellemzi, hogy a teszt milyen megbízhatóan detektálja a betegség fennállását."(Fidy & Makara, 2005)

2.1.3. Fontos megjegyzés

Fontos megjegyezni, hogy az ilyen tesztek célja nem az, hogy a tesztalanynak minél nagyobb valószínűséggel jelezzen pozitívat, ha ténylegesen fertőzött. A tesztek

legfőbb célja a hamis negatívak csökkentése, tehát azon esély csökkentése, melynél a negatív teszteredmény esetén valójában fertőzött a tesztalany, mivel a cél a fertőzők izolálása, nehogy tévedésből el legyenek engedve.

2.2. A tesztek

2.2.1. A PCR teszt

A PCR teszt során a tesztalany garatjából és orr nyálkahártyájából vesznek mintát, az ott található sejtekből. A mintát egy pálcika segítségével szerzik be úgy, hogy azt a garatfalhoz és az orrjárat hátsó részéhez nyomják. A mintavétel után nagyjából 1-3 nappal később derül ki az eredménye a vizsgálatnak. A vizsgálat során az alanytól vett mintából a vírus örökítőanyagát próbálják meg kimutatni. A PCR teszt a legspecifikusabb és legérzékenyebb módszer. Mivel több helyen is végeznek PCR vizsgálatokat, különböző módokon, így a pontos szenzitivitást és specifitást nem tudjuk megmondani, csak annyit, hogy körülbelül a teszt specifitása 85-93% között, a szenzitivitása pedig 95-96% között van. Az egyszerűség kedvéért a továbbiakban számoljunk 89%-os specifitással és 95%-os szenzitivitással. (Bilecz, 2020)

2.2.2. Rapid antitest teszt („gyorsteszt")

A gyorsteszt legnagyobb előnye, ahogy azt neve is sugallja, hogy gyorsan végezhető. A tesztalany vérmintáját vizsgálják, melyben a koronavírus specifikus ellenanyagait keresik. Az eredmény 15 percen belül elkészül. Rengeteg fajta gyorsteszt létezik, viszont az egyszerűség kedvéért egy olyannal fogunk foglalkozni, amelynek a specificitása 99% és a szenzitivitása 60%. (Dimenzió-Med Egészségközpont, 2020)

2.2.3. Rapid antigén teszt

A rapid antigén teszt során a tesztalanytól légúti mintát vesznek. Elsősorban a nagy valószínűséggel fertőzött embereknél szokták használni, ugyanis ők a vírus S-proteinjét a legutak segítségével ürítik ki. A teszt elvégzése itt is elég gyors, 15 percen belül kész lehet az eredmény. A teszt specificitását itt is 99%-nak fogjuk mondani és az érzékenysége pedig 84-97,6% közötti értékeket vehet fel, így mi a 91%-ot fogjuk használni. (Centers for Disease Control and Prevention, 2020)
Megjegyzés: Eredetileg a gyors, illetve a rapid antigén tesztet is úgy hirdették, hogy 100%-os a specificitásuk. Azonban 100 helyett csak 99%-kal számolunk, ugyanis a Bayes-statisztika egyik alapelve, hogy minden értéket a (0,1)-ból választunk.

2.2.4. Megjegyzés a tesztek paramétereiről

A PCR teszt esetében az érzékenység értéke magasabb a specificitás értékénél. A másik két teszt esetében ez pont fordítva van. Ez megfigyelhető más vírusokra kifejlesztett tesztek esetében is, például a HIV vírus esetében is (Peabody, 2019).
Ez a különbség azért van, mivel a magasabb érzékenység következménye több fals pozitív eredmény, és hasonlóan a magasabb specificitás következménye a több fals

negatív eredmény. A két érték közötti megfelelő egyensúly elérése a tesztek célja. Ezért az egyes teszteket más-más alkalmakkor használják. A HIV vírus tesztelése esetében két tesztes megoldást alkalmaznak, ugyanis végrehajtanak egy magas érzékenységű tesztet és egy magas specificitású tesztet is. Az elsőt azért, hogy minél több fals negatívat szűrjenek ki, majd a második során minél több fals pozitívat (Peabody, 2019).
A koronavírus esetében a PCR tesztet javasolják egyéni, valamint kisebb közösségek tesztelésére. A rapid antigén tesztet olyan esetben javasolják, amikor a pozitív eredmény valószínűsége várhatóan magas, ennek megerősítésére. A gyorstesztet is hasonló felhasználásra javasolják, mint a rapid antigén tesztet (Synlab, 2020).

2.3. A tesztek hatékonysága

Ahhoz, hogy megállapítsuk, hogy milyen hatékonyak az általunk vizsgált tesztek, meg kell néznünk, hogy vajon hány fertőzött ember lehet a lakosság körében. A valóban fertőzött emberek számát nem lehet megmondani, csak a beigazolódott fertőzöttekét. Úgy becsülik, hogy a valóban fertőzöttek akár több mint tízszer annyian is lehetnek, mint a regisztrált fertőzöttek. Október 8-án az aktív fertőzöttek száma Magyarországon 23 088 fő volt (Magyarország Kormánya, 2020).
Az ország lakossága (2019-es adatok alapján) 9 772 756 fő.(Központi Statisztikai Hivatal, 2020) Ezek alapján a lakosság 23 088 / 9 772 756 = 0,24% -a regisztrált fertőzött. Tegyük fel, hogy a valódi fertőzöttek száma ennek tízszerese.(Havers, Reed, Lim, Montgomery, Klena, Hall... Thornburg, 2020) Ebben az esetben a lakosság ˜ 2,5%-a fertőzött. A továbbiakban meg fogjuk vizsgálni a tesztek hatékonyságát egy 10 000 fős populáción.

2.3.1. A PCR teszt

Tegyük fel, hogy 10 000 PCR tesztet elvégzünk véletlenszerűen a lakosság körében. A 2,5%-kal számolva 250 fő fertőzött a 10 000 főből. Mivel tudjuk, hogy a teszt érzékenysége 95%, így könnyen ki tudjuk számolni, hogy a $0{,}95 \cdot 250 = 238$ főnél lesz pozitív a teszt és 12-nél pedig negatív. A 89%-os specificitásból pedig könnyen kiszámolható, hogy $9750 \cdot 0{,}89 = 8678$ főnél fog a teszt negatívat mutatni, és 1072 főnél pedig pozitívat. Ezt az alábbi ábra is szemlélteti:

		Valóság	
		Pozitív	Negatív
Teszt eredmény	Pozitív	238	1072
	Negatív	12	8678

2.3.2. Rapid antitest teszt („gyorsteszt”)

Az előző teszthez hasonlóképpen 2,5%-os fertőzöttséggel és 10 000 fővel, valamint a teszt 99%-os specificitásával és 60%-os érzékenységével számolva a teszt a 250-ből 150 főnél fog pozitívat mutatni, és a 9750 főből pedig 9653-nál fog negatívat mutatni.

		Valóság	
		Pozitív	Negatív
Teszt eredmény	Pozitív	150	97
	Negatív	100	9653

2.3.3. Rapid antigén teszt

Szintén 2,5%-os fertőzöttséggel és 10 000 fővel, valamint a teszt 99%-os specificitásával és 91%-os érzékenységével számoltunk. Ebben az esetben a teszt 250-ből 228 főnél mutat pozitívat, és 9653 főnél a 9750-ből negatívat.

		Valóság	
		Pozitív	Negatív
Teszt eredmény	Pozitív	228	97
	Negatív	22	9653

2.4. A tesztek megbízhatósága

Ezek alapján szeretnénk megtudni, hogy mekkora az esélye, hogy valaki ténylegesen fertőzött, ha pozítivak az egyes tesztjei?
A Wickmann (1995) Bayes-statisztika című könyv jelölései alapján, jelöljük θ-val a tényleges egészségügyi állapotot (tehát θ_1 = fertőzött és θ_2 = nem fertőzött), és X pedig legyen a teszteredmény ($x = 0$ negatív és $x = 1$ pozitív).
A jelölések alapján a kíváncsiak vagyunk arra, hogy:

$$P(\theta_1|x = 1) = ?$$

Nézzük meg ezt az egyes tesztekre.

2.4.1. A PCR teszt

A feltételes valószínűség és Bayes-tétel alapján:

$$P(\theta_1|x = 1) = \frac{P(\theta_1)P(x = 1|\theta_1)}{P(\theta_1)P(x = 1|\theta_1) + P(\theta_2)P(x = 1|\theta_2)}$$

A következő adatokat tudjuk:

$$P(\theta_1) = 0.025$$
$$P(\theta_2) = 0.975$$

$$P(x = 1|\theta_1) = 0.95 \text{ (érzékenység)}$$

$P(x = 1|\theta_2)$-t pedig a feltételes valószínűség segítségével ki tudjuk számolni:

$$P(x = 1|\theta_2) = \frac{P(x = 1 \cap \theta_2)}{P(\theta_2)} = 0.11$$

A fenti adatokat pedig a Bayes-tételbe behelyettesítve a következőt kapjuk:

$$P(\theta_1|x = 1) = \frac{0.025 \cdot 0.95}{0.025 \cdot 0.95 + 0.975 \cdot 0.11} = 0.1851$$

Láthatjuk, hogy ha a teszt pozitív eredményt mutat, akkor is csak 19% esély van arra, hogy ténylegesen fertőzöttek vagyunk. Ez elsőre meglepőnek tűnik, viszont ha jobban belegondolunk ez az ára annak, hogy ha negatívat mutat a teszt, akkor 99% eséllyel tényleg ne legyünk fertőzöttek.

2.4.2. Rapid antitest teszt („gyorsteszt")

Az előzőhöz hasonló számolással:

$$P(x = 1|\theta_1) = 0.6$$

$$P(x = 1|\theta_2) = 0.01$$

$$P(\theta_1|x = 1) = 0.6073$$

2.4.3. Rapid antigén teszt

Az előzőhöz hasonló számolással:

$$P(x = 1|\theta_1) = 0.91$$

$$P(x = 1|\theta_2) = 0.01$$

$$P(\theta_1|x = 1) = 0.7015$$

2.5. Több teszt elvégzése

Ezek után felmerülhet a kérdés, hogy mi van akkor, ha egymás után készítünk el több tesztet, akkor az vajon megbízhatóbb eredményt fog adni? Nézzük meg, hogy javít-e a megbízhatóságon az, hogyha a gyorsteszt után a másik két fajta tesztet is elvégezzük. Tegyük fel, hogy a gyorstesztünk eredménye pozitív lett. Mekkora esély van ekkor, hogyha elvégzünk egy másik tesztet, akkor az is pozitív lesz?

2.5.1. Gyorsteszt után PCR teszt

A Bayes-i gondolkodásnak megfelelően fogunk eljárni. A gyorsteszt előtt a priori eloszlásunk a következő volt:

$$P_1(\theta_1) = 0.025$$

$$P_1(\theta_2) = 0.975$$

Mivel a gyorsteszt pozitív lett, így az ebből kapott poszteriori eloszlást fogjuk a következő kísérlet priori eloszlásának használni:

$$P_2(\theta_1) = 0.6073$$

$$P_2(\theta_2) = 0.3927$$

Ha ezekkel az adatokkal végezzük el a kísérletet, akkor a következő poszteriori eloszlást fogjuk kapni:

$$P_2(\theta_1|x = 1) = \frac{P_2(\theta_1)P_2(x = 1|\theta_1)}{P_2(\theta_1)P_2(x = 1|\theta_1) + P_2(\theta_2)P_2(x = 1|\theta_2)} = \frac{0.6073 \cdot 0.95}{0.6073 \cdot 0.95 + 0.3927 \cdot 0.11} = 0.9303$$

Az eredmény 93,03%, ami elég magasnak mondható a 60%-hoz és az 19%-hoz képest is. Tehát ha két tesztet végzünk, akkor az javítani fogja az eredményt. Nézzük meg hasonlóan, hogyha egy gyorstesztet és egy rapid antigén tesztet végzünk el.

2.5.2. Gyorsteszt után rapid antigén teszt

A gyorsteszt utáni poszteriori eloszlást most priori eloszlásnak használva a következő lesz a kísérlet:

$$P_2(\theta_1|x=1) = \frac{P_2(\theta_1)P_2(x=1|\theta_1)}{P_2(\theta_1)P_2(x=1|\theta_1) + P_2(\theta_2)P_2(x=1|\theta_2)} = \frac{0.6073 \cdot 0.91}{0.0673 \cdot 0.91 + 0.3927 \cdot 0.01} = 0.9403$$

Láthatjuk, hogy az eredmény itt is javult.

2.5.3. PCR teszt után rapid antigén teszt

Nézzük meg, hogy a PCR teszt eredményén tudunk-e javítani a Rapid Antigén teszt segítségével.
A poszteriori eloszlásunk most a következő:

$$P_2(\theta_1) = 0.19$$
$$P_2(\theta_2) = 0.81$$

Végezzük el a kísérletet most a rapid antigén teszttel:

$$P_2(\theta_1|x=1) = \frac{P_2(\theta_1)P_2(x=1|\theta_1)}{P_2(\theta_1)P_2(x=1|\theta_1) + P_2(\theta_2)P_2(x=1|\theta_2)} = \frac{0.19 \cdot 0.91}{0.19 \cdot 0.91 + 0.81 \cdot 0.01} = 0.9552$$

Láthatjuk, hogy jelentős növekedést értünk el a két teszt egymás utáni alkalmazásával. Azaz két különböző egymásutáni pozitív teszt után már tényleg nagy valószínűséggel mondható, hogy fertőzött vagyok. De jegyezzük meg, hogy nem ez a teszt elsődleges pozitívuma, hogy pozitív esetben nagy eséllyel tényleg fertőzött vagyok-e, hanem az, hogy kiszűrje potenciális veszélyforrásokat.

2.6. Esetszám változtatás

A számolás elején az október 8-i adatokkal számoltunk. A tényleges fertőzöttek számát pedig a tesztek által megállapított fertőzöttek számának tízszeresével becsültük. Nézzük meg, hogyan változik a tesztek megbízhatósága abban az esetben, ha a tényleges fertőzöttek számát az október 8-i állapot tesztek által megállapított fertőzöttei adnák csak ki. Tehát ebben az esetben a ténylegesen fertőzöttek száma a lakosság 0.24%-a.

2.6.1. Ha kevesebb a ténylegesen fertőzött

Az előzőekben látott számolások alapján nézzük meg mekkora az esélye, hogy fertőzöttek vagyunk, ha az egyes tesztek pozitívat mutatnak:

a) PCR teszt esetében: $P(\theta_1|x = 1) = 0.0214$

b) Gyorsteszt esetében: $P(\theta_1|x = 1) = 0.1304$

c) Rapid antigén teszt esetében: $P(\theta_1|x = 1) = 0.1869$

2.6.2. Ha több a ténylegesen fertőzött

Érdemes ezek után megvizsgálni azt is, hogy mi történik abban az esetben, ha növeljük a ténylegesen fertőzöttek számát. November 10-én az aktív fertőzöttek száma hazánkban 88 737.(Magyarország Kormánya 2020) Ez a lakosság 0,91%-a. A valódi fertőzöttek számához, ha megint a tízszeres becslést alkalmazzuk, akkor a lakosság 9,1%-át mondhatjuk fertőzöttnek. Nézzük meg, hogy alakulnak az eredmények ezekkel az adatokkal számolva.

a) PCR teszt esetében: $P(\theta_1|x = 1) = 0.4638$

b) Gyorsteszt esetében: $P(\theta_1|x = 1) = 0.8571$

c) Rapid antigén teszt esetében: $P(\theta_1|x = 1) = 0.901$

2.7. Kérdőív a koronavírus tesztek megbízhatóságáról

2.7.1. Kérdések

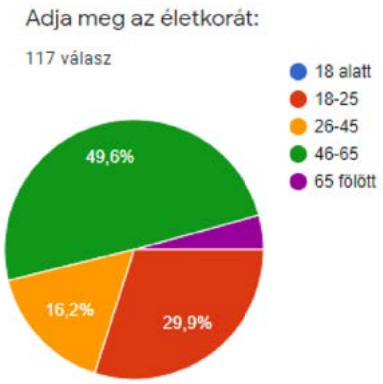
A szakdolgozatomhoz kapcsolódóan készítettem egy kérdőívet, az eddig tárgyalt témával kapcsolatban. A kérdőívet 2020.10.28. és 2020.11.04. között lehetett kitölteni. Ez idő alatt 117 válasz érkezett. A kitöltők sajnos nem reprezentálják a teljes lakosságot, így az eredményeket óvatossággal kell kezelni. A kérdőív során az adatfelvételen kívül hat kérdést tettem fel, melyek a következők:

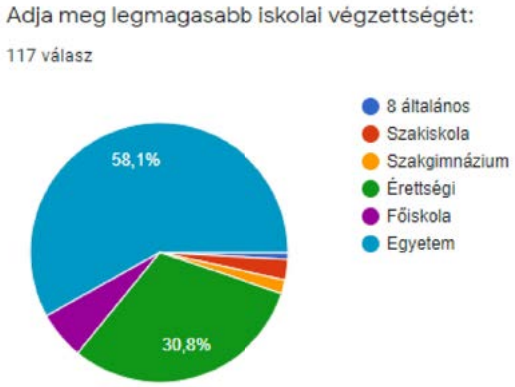
- Mit gondol a PCR teszt a ténylegesen fertőzöttek hány százalékát mutatja pozitívnak?
- Mit gondol a gyorsteszt a ténylegesen fertőzöttek hány százalékát mutatja pozitívnak?
- Mit gondol a rapid antigén teszt a ténylegesen fertőzöttek hány százalékát mutatja pozitívnak?
- A PCR tesztnél ha pozitív eredményt kap, mit gondol mekkora a valószínűsége, hogy tényleg fertőzött?
- A gyorstesztnél ha pozitív eredményt kap, mit gondol mekkora a valószínűsége, hogy tényleg fertőzött?
- A rapid antigén tesztnél ha pozitív eredményt kap, mit gondol mekkora a valószínűsége, hogy tényleg fertőzött?

2.7.2. A kitöltők adatai

A kitöltőktől a következő adatokat kértem el:

- Életkor
- Legmagasabb iskolai végzettség
- Foglalkozás
- Legmagasabb fokú matematikai tanulmány





A fenti ábrák alapján jól látható, hogy a kitöltők nagy része 46-65 év közötti, valamint a kitöltők többsége egyetemi szintű matematikai tanulmánnyal rendelkezik. Ezek alapján érdemes megvizsgálni, hogy vajon mennyire különböző eredménnyel töltötték ki a matematikában kevésbé jártas és a matematikában jártas kitöltők a kérdőívet. Ez alapján nevezzük a továbbiakban *egyes* csoportnak, akik legfeljebb közép szintű érettségi szintű matematika végzettséggel rendelkeznek, és *kettes* csoportnak, akik emelt szintű érettségivel, vagy egyetemi szintű matematikai végzettséggel rendelkeznek.

2.7.3. Az eredmények

Az első három kérdés az egyes tesztek érzékenységére utalt. A PCR teszt esetében én 95%-kal számoltam. Az egyes csoport által írt válaszok átlaga 70,6%, a kettes csoport által írt válaszok átlaga pedig 81% lett.

A gyorsteszt esetében én 60%-os érzékenységgel számoltam, az egyes csoport átlaga 61,7%, a kettes csoport átlaga pedig 67,4% lett.

A rapid antigén tesztnél pedig 91%-os érzékenységgel számoltam, az egyes csoport átlaga pedig 57,8% lett, a kettes csoporté pedig 63,1%.

Ezután következtek azok a kérdések, melyeknél azt kérdeztem, hogy ha pozitív eredményt mutat a teszt, akkor mekkora valószínűséggel vagyunk tényleg fertőzöttek.

A PCR teszt esetében én 18,5%-ot számoltam ki. Az egyes csoport 74,5%-ot, a kettes pedig 80,75%-ot írt.

A gyorsteszt esetében 60,7%-ot számoltam, az egyes csoport 60,5%-ot írt, a kettes pedig 69,9%-ot.

70,2% lett a rapid antigén teszt esetében az eredményem, míg az egyes csoport itt 60,7%-ot, a kettes pedig 65,6%-ot írt.

2.7.4. Az eredmények kiértékelése

A beérkezett válaszok alapján úgy tűnik, hogy a kettes csoport mind a hat kérdésnél magasabb százalékokat írt. A legjobban kitűnő eredmény a PCR tesztnél alakult ki. Az általam a Bayes-tétel segítségével kiszámolt 18,5%-kal szemben, a válaszadók 74,5%, illetve 80,75%-ot írtak be. Ez a nagy különbség valószínűleg azért van, mivel mostanság az van a köztudatban, hogy a PCR teszt a legmegbízhatóbb. Ez alapján és a józan ész alapján az emberek azt mondhatják, hogy akkor biztos az a teszt fogja megmondani a legtöbb esetben azt is, hogy ténylegesen fertőzött-e az illető, ha pozitív a teszt. Ezt a válaszok is tükrözik, ugyanis mindkét csoportnál a PCR teszt esetén adták a legnagyobb értékeket. A kiszámolt 18,5% pedig megmutatja, hogy nem mindig érdemes a józan eszünkre támaszkodni.

2.7.5. A többi teszt

Ezek után jogosan merülhet fel bennünk a kérdés, hogy akkor a másik két teszt miért hoz jobb eredményt, mint a PCR teszt, ha a PCR teszt a legmegbízhatóbb. Első sorban erre a válasz az, hogy a PCR teszt nem a legmegbízhatóbb, hanem a legérzékenyebb teszt. Másrészről pedig ha megvizsgáljuk a gyorsteszt, valamint a rapid antigént teszt paramétereit, akkor láthatjuk, hogy a specificitást tekintve mindkettő jobb, mint a PCR teszt. Az adatok hitelessége miatt viszont kétségesek a kiszámolt eredmények. A legtöbb az interneten megvásárolható gyorsteszten nem találhatóak meg ezek a paraméterek, vagy pedig az általunk számolt adatokhoz hasonlóakról szólnak, vagy nem rendelkeznek diagnosztikai minősítéssel. Ezek alapján előfordulhat, hogy a valóságban a gyorsan elvégezhető tesztek (gyorsteszt és rapid antigén teszt) paramétereinek értékei alacsonyabbak. Ezek alapján természetesen a kiszámolt értékek is csökkennének, de mivel konkrét adatokkal jelenleg nem tudunk számolni, ezért a fellelhető adatokból számoltam.

2.8. Konklúzió

Mivel jelenleg is tart a járvány, így sok bizonytalansággal kell számolnunk a tesztekkel kapcsolatban. A valós adatokat pontosan nem tudjuk, és a tesztek teljesítményéről is csak körülbelüli fogalmunk van. Ennek ellenére állítható, hogy a PCR teszt a legérzékenyebb az összes közül, és ha valakinél felmerül a fertőzés gyanúja, akkor érdemes elvégezni nála azt. A kiszámolt eredmények alapján láthattuk, hogy nem szabad mindig pusztán a megérzésünkre támaszkodni. A józan ész és a matematika szerint kell eljárnunk, ha a valósághoz közelebbi eredményt szeretnénk kapni.

3. fejezet

A hanglemez kereskedő

3.1. Alapprobléma

A hanglemez kereskedőnek felajánlanak 1000 db hanglemezt, 10 euró/db áron. (Feltételezhetjük, hogy ez a raktár és a munkaköltségi árat is tartalmazza). A hanglemezeket 15 euró/db áron tudja eladni. A hibás hanglemezek aránya ismeretlen. A kereskedőnek két választási lehetősége van. Az első, hogy átvegye a szállítmányt, a második, hogy F szállítási vállalat ajánlatát elfogadja. F szállítót a kereskedő már rég óta ismeri, és tudja, hogy az ő szállítmányában nem található selejt. Ha a kereskedő a második lehetőséget választja, akkor 2000 euró biztos nyereséggel számolhat. A döntés előtt az ismeretlen szállító árujából 75 lemezt megvizsgálhat a kereskedő. Vajon melyik döntés fog nagyobb nyereséget hozni?

3.2. A Bayes-elvről általánosan

A Wickmann (1995) könyvből idézve:

3.2.1. A Bayes elv diszkrét esetben

„Adott[...]:

1. az $\mathcal{A} = \{a_i | 1 \leq i \leq \mathcal{I}\}$ halmaz, amely az adott szituációban lehetséges, egymást kizáró a_i cselekvéseket tartalmazza; $\mathcal{A}$ az ún. cselekvés- vagy döntéstér; $|\mathcal{A}| = \mathcal{I}$;

2. az $\Omega = \{E_j | 1 \leq i \leq \mathcal{J}\}$ halmaz, amely $\mathcal{J}$ darab egymást kizáró, lehetséges E_j eseményt és egy $P(E_j)$ valószínűségeloszlást tartalmaz;

3. a $g_i(E_j)$ nyereségfüggvények úgy, hogy az a_i-hez tartozó εg_i érték $\varepsilon g = \sum_{j=1}^{J} g(E_j)P(E_j)$, ahol $\sum_{j=1}^{J} P(E_j) = 1$ alapján adott." (Wickmann 1995, p.58)

Azt az $a_* \in \mathcal{A}$ cselekvést kell választanom, amelyre teljesül (Wickmann 1995, p.58):

$$\varepsilon g_* = max\{\varepsilon g_i\}.$$

3.2.2. A Bayes tétel folytonos esetben

A folytonos θ eseménytérre vonatkozó Bayes-tétel az $f(\theta|x)$ poszteriori sűrűségfüggvényre.(Wickmann 1995, p.67)

$$f(\theta|x) = \frac{f(\theta)P(x|\theta)}{\int_\theta f(\theta)P(x|\theta)d\theta}$$

3.2.3. A Bayes elv folytonos esetben

A diszkrét esethez hasonlóan itt is maximális nyereségű a_*-ot kell választanunk. Azonban itt $\mathcal{J}$ darab szorzat összege helyett itt az összes lehetséges θ-t kell vennünk, melyet a θ-n való integrálással tudunk kifejezni. ($\theta = \{\theta_j | j \in \{1, ..., \mathcal{J}\}\}$ diszkrét esetben, és most $\theta = [0, 1]$ és p egy θ beli valós szám.) Így (Wickmann 1995, p.67):

$$\varepsilon g_i = \int_\theta g_i(\theta) f(\theta|x) d\theta$$

3.3. A feladat megoldása

A feladat megoldása a Wickmann (1995) könyv alapján. Legyen p a selejtek aránya az ismeretlen szállító szállítmányában. Legyen g_1 és g_2 a nyereségfüggvények az egyes esetekben.

$$g_1(p) = 1000 \cdot 15 \cdot (1 - p) - 10000 = 5000 - 15000p$$
$$g_2(p) = 2000$$

A Bayes-elv használatához a következő feltevésekre van szükségünk:

- Tegyük fel, hogy az 1000 db hanglemez között a selejtes hanglemezek egyenletes eloszlást követnek. Ekkor $f(p) = 1$.
- Mivel véletlenszerűen választunk ki 75 db lemezt a vizsgálat céljára, így ez egy Bernouilli-kísérlet lesz. Tehát a megvizsgált n = 75 lemez között k db lesz selejtes.

A folytonos esetre vonatkozó Bayes-tétel segítségével meg tudjuk határozni a poszteriori sűrűségfüggvényt.

$$f(p|k) = \frac{f(p) \cdot \binom{n}{k} p^k (1-p)^{n-k}}{\int_0^1 f(p) \cdot \binom{n}{k} p^k (1-p)^{n-k} dp} = \frac{p^k (1-p)^{n-k}}{\int_0^1 p^k (1-p)^{n-k} dp}$$

A $\int_0^1 x^a (1-x)^b dx = \frac{a! \cdot b!}{(a+b+1)!}$ összefüggés segítségével a kifejezés átalakítható a következőre:

$$f(p|k) = \frac{(n+1)!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Az eddigiek alapján a nyereségfüggvényünk pedig a következő alakot veszi fel:

$$\varepsilon g_i = \int_0^1 g_i(p) f(p|k) dp$$

Mivel nem tudjuk k mekkora értéket vesz fel pontosan, így a megfelelő döntés érdekében meg kell határoznunk azt a c küszöbszámot, ami felett már a második választási lehetőség fog nagyobb nyereséget hozni.
Helyettesítsük be $k = c$-t!

$$f(p|c) = \frac{(n+1)!}{c!(n-c)!}p^c(1-p)^{n-c}$$

Ekkor a nyereségfüggvény a következő:

$$\varepsilon g_1 = \int_0^1 (5000 - 15000p)\frac{(n+1)!}{c!(n-c)!}p^c(1-p)^{n-c}dp$$

$$= \int_0^1 5000 \cdot \frac{(n+1)!}{c!(n-c)!}p^c(1-p)^{n-c}dp - \int_0^1 15000 \cdot \frac{(n+1)!}{c!(n-c)!}p^{c+1}(1-p)^{n-c}dp$$

$$= 5000 \cdot \frac{(n+1)!}{c!(n-c)!}\int_0^1 p^c(1-p)^{n-c}dp - 15000 \cdot \frac{(n+1)!}{c!(n-c)!}\int_0^1 p^{c+1}(1-p)^{n-c}dp$$

$$= 5000 \cdot \frac{(n+1)!}{c!(n-c)!} \cdot \frac{c! \cdot (n-c)!}{(c+n-c+1)!} - 15000 \cdot \frac{(n+1)!}{c!(n-c)!} \cdot \frac{(c+1)! \cdot (n-c)!}{(c+1+n-c+1)!}$$

$$\varepsilon g_1 = 5000 - 15000 \cdot \frac{c+1}{n+2}$$

Tegyük egyenlővé εg_1 és εg_2-t.

$$\varepsilon g_1 = \varepsilon g_2$$

$$5000 - 15000 \cdot \frac{c+1}{n+2} = 2000$$

$$\frac{c+1}{n+2} = \frac{1}{5}$$

$n = 75$-re

$$c = 14,4$$

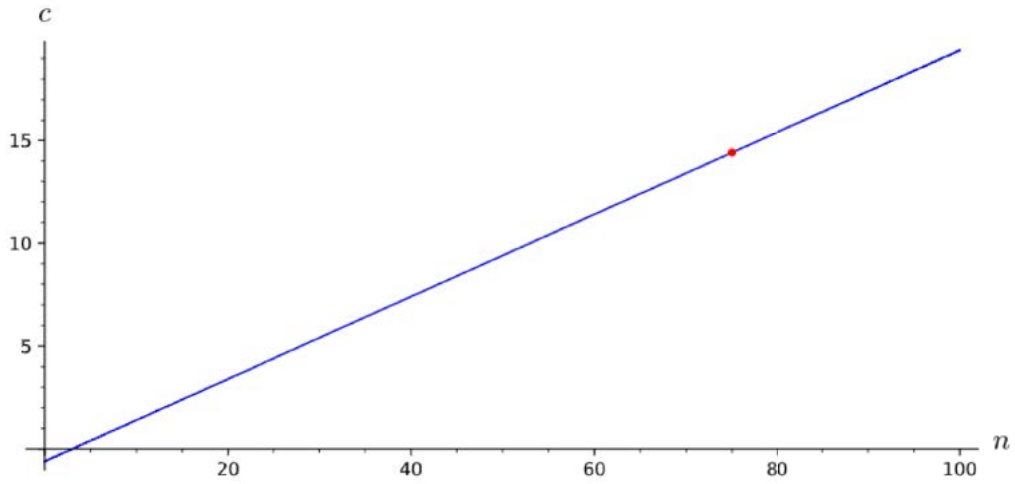
Ez azt jelenti, hogy $k \leq 14$ esetén érdemes a_1 szállítmányt választani, egyébként a_2-t.

3.4. n változtatása

Érdemes megvizsgálni mi történik akkor ha változtatjuk a lemezek mintavételének számát. A feladatban nem szerepel, hogy a mintavétel költséggel járna-e.

3.4.1. Ha ingyenes a mintavétel

Ha semmilyen költséggel nem jár a mintavétel, akkor azt ahogyan az alábbi ábrán is láthatjuk, minél nagyobb az n értékünk, annál nagyobb lesz a c érték is, tehát nagyobb mintavétellel pontosabb értéket kapunk c meghatározásakor. A hányadosuk pedig közelíteni fog p-hez.



Viszont ez a valóságot nem tükrözi, ugyanis a mintavétel a valóságban általában nem költségmentes.

3.4.2. Ha a mintavétel nem ingyenes

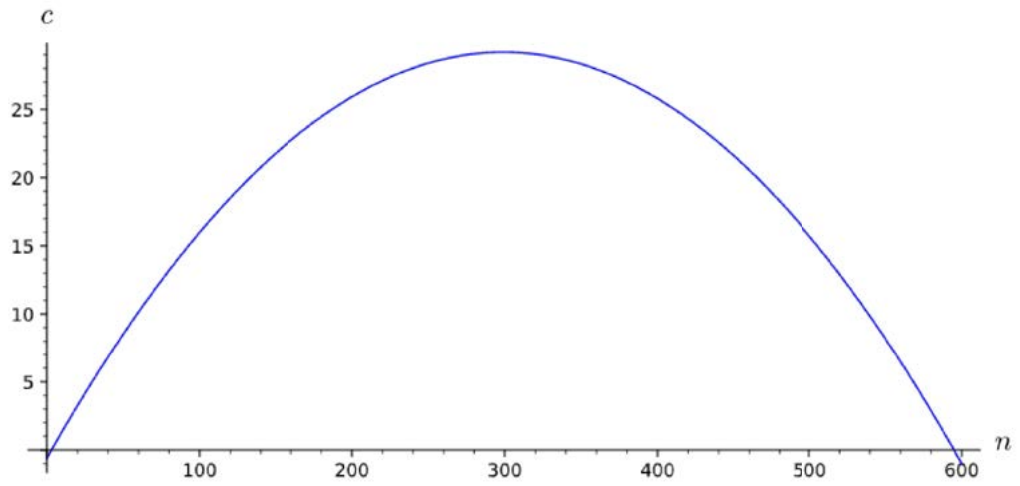
Legyen l a költsége minden egyes mintavételnek.
Ez által az eredeti nyereségfüggvényünk a következőképpen módosul:

$$g_1(p,n,l) = 5000 - 15000p - n \cdot l$$
$$g_2(p,n,l) = 2000$$

$$\varepsilon g_1(p,n,l) = 5000 - 15000 \cdot \frac{c+1}{n+2} - n \cdot l$$

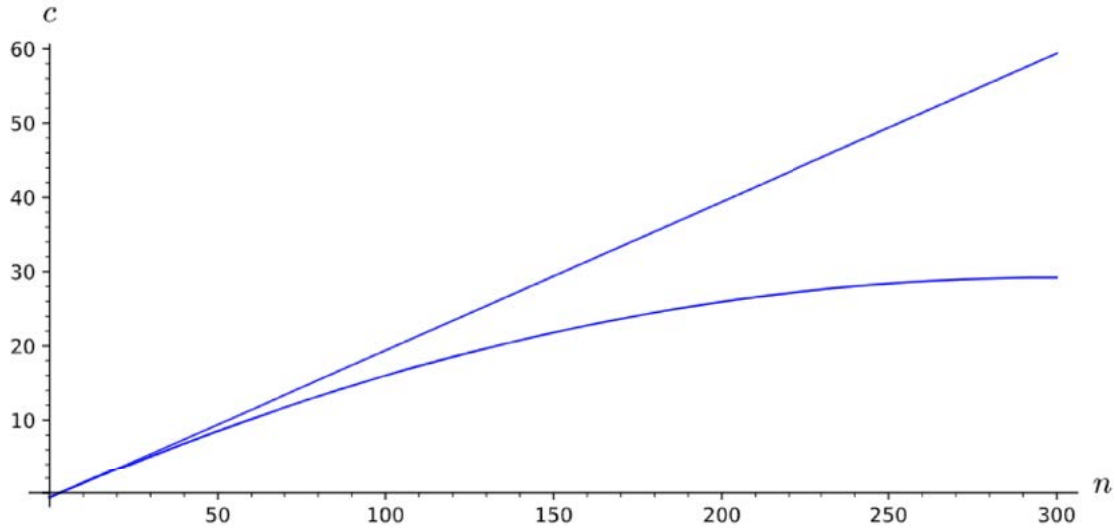
Tegyük fel, hogy $l = 5$. Hasonlítsuk össze a két nyereségfüggvényt, hogy mikor lesz $\varepsilon g_1 \geq \varepsilon g_2$

$$5000 - 15000 \cdot \frac{c+1}{n+2} - n \cdot 5 \geq 2000$$
$$3000 \geq 15000 \cdot \frac{c+1}{n+2} + 5n$$
$$3000n + 6000 \geq 15000c + 15000 + 5n^2 + 10n$$
$$-15000c \geq 9000 + 5n^2 - 2990n$$
$$c \leq -\frac{9}{15} - \frac{n^2}{3000} + \frac{2990n}{15000}$$

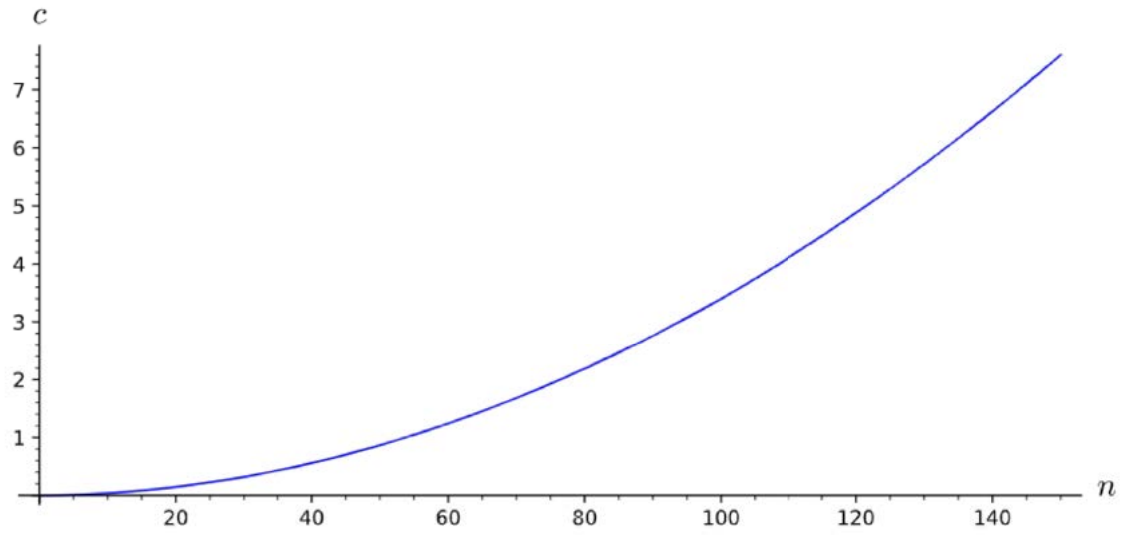


A fenti ábra alapján látható, hogy mely tartományban lesznek a számunkra nyereséges kombinációk. A görbe alatti területen lesz több mint 2000 a nyereségünk. A függvény a maximumát 299-nél veszi fel. Ebből azt a következtetés vonhatjuk le, hogy n-et nem érdemes 299 fölé növelnünk, ugyanis akkortól c értéke elkezd csökkenni. Ez azért érdekes, mivel feltételezhetjük, hogy több selejtes minta lesz 299-nél több mintában, mint 299-nél kevesebb mintában. Viszont a maximális c érték $n = 299$ fölött már csökken, így ha $n = 299$-nél még hezitálunk a nagyobb mintavétellel, akkor tudhatjuk, hogy már biztosan nem érdemes tovább menni. Így a tartományunk leszűkül $[1, 299]$-re.

Ezután érdemes megnéznünk, milyen viszonyban van az ingyenes mintavétel által kapott függvény és a nem ingyenes mintavétel által kapott függvény.



Az ábráról könnyen leolvasható, hogy a két függvény közötti különbség n növelésével egyre csak növekedni fog. Ez számunkra azért érdekes, mivel ha a két függvény "közel" van egymáshoz, akkor a mintavételi költség kicsi, és minél távolabb vannak egymástól, annál nagyobb ez a költség. Ez alapján azt mondhatjuk, hogy n-et addig érdemes növelni csak, amíg a mintavételi költség kicsi, tehát a két függvény "közel" van egymáshoz. De pontosan mit értünk "kicsi" alatt? Ehhez nézzük meg a két függvény különbségét. Vonjuk ki egymásból a két függvényt.



Az így kapott különbségfüggvény megmutatja, hogy melyik n értékhez mekkora c-beli különbség lesz a két függvény között. Azt tudjuk, hogy ha növeljük n-et, akkor a mintavételi költség miatt csak a kisebb c értékek lesznek már számunkra elfogadhatóak. Ezt az előző ábrán is láthatjuk. Mivel számunkra csak a kisebb c értékek lesznek elfogadhatóak, így csökkenni fog a valószínűsége annak, hogy a szállítmányt át tudjuk venni, mivel kevesebb selejtet engedünk meg csak ekkor. Így nem érdemes túl nagy n-et néznünk, mert akkor c nagyon le fog csökkenni, azonban valamennyi mintát muszáj vennünk, hogy megtudhassunk valamit a szállítmányról. Így a különbségfüggvény segíthet abban, hogy eldöntsük, hogy mekkora maximális c csökkenésig vagyunk hajlandóak elmenni.

Ha azt mondjuk, hogy a maximális c érték csak 1-gyel csökkenhet, akkor a különbségfüggvénybe $c = 1$-et helyettesítve megkapjuk, hogy akkor $n = 53$-ig mehetünk. Ebben az esetben információt is kapunk a szállítmányról és a mintavétel költsége még nem is befolyásolja nagyon a végeredményt. Ennek alapján egy 50 körüli mintavétel jó döntésnek tűnik.

4. fejezet

A gyümölcskereskedő példája

4.1. Alapprobléma

Ez a feladat az előző fejezetben tárgyalt hanglemez kereskedő példájának folytatása. Az eddigiek során a Bayes-elv alapján mindig a legjobb $a_* \in \mathcal{A}$ cselekvést akartuk választani, ahol $\mathcal{A}$ véges sok $\mathcal{I}$ esemény halmaza volt. A hanglemez kereskedő példájában $\mathcal{I} = 2$ volt, viszont most vizsgáljuk meg azt az esetet, ha a lehetséges események halmaza nem diszkrét, hanem folytonos. Ez alapján a vizsgált probléma a következő: Egy gyümölcskereskedő azon gondolkodik, hogy mennyi narancsot rendeljen a következő hétre. A narancs romlandó, így egy hét után már nem lehet eladni. Jelöljük a-val a kereskedő által megvásárolt készletet, x-szel pedig a keresletet. Tegyük fel, hogy a felvásárlási és eladási ár aránya 1:4. Ekkor a következő folytonos nyereségfüggvényt kapjuk:

$$g(a,x) = \begin{cases} 4x - a, & \text{ha } x \leq a \\ 3a, & \text{ha } x \geq a \end{cases}$$

Határozzuk meg azt a legjobb a_* értéket, amire a kereskedő az adott keresletkínálathoz mérten a legnagyobb nyereségre tesz szert (Wickmann 1995).

4.2. Feladatmegoldás

4.2.1. Becslés a keresletre

A feladat megoldása a Wickmann (1995) könyv alapján. A gyümölcskereskedő ezenkívül még egy becslést is adott a keresletre. Azt mondta, hogy a közepes kereslet 1000, és a közepestől való pozitív és negatív eltérésnek azonos a valószínűsége, nagyjából nyolc hét alatt egyszer fordul elő, hogy legfeljebb 500-at tud eladni. Tegyük fel, hogy f (kereslet) most normális eloszlást követ. Így kiszámolhatjuk a sűrűségfüggvényt: $f_N(x|\mu,\sigma)$. A közepes kereslet 1000, így legyen $\mu = 1000$. σ pedig a kereskedő elmondása alapján megállapítható a tapasztalatok alapján. A legfeljebb 500-as kereslet így nagyjából $\frac{1}{8}$. A standardizált normális eloszlás alapján, az inverz függvény segítségével megkapjuk, hogy $z(F = 0.125) = -1.15$. Mivel $\frac{x-\mu}{\sigma} = z$, így $\sigma = \frac{500-1000}{-1.15} \approx 435$. Tehát $\sigma = 435$-tel számolhatunk.

4.2.2. Veszteségfüggvény

Az előző alfejezetben meghatározott nyereségfüggvényünk átalakítható egy vele ekvivalens nyereségfüggvénnyé:

$$g_{ekv} = g(a, x) - g'(x)$$

Ahol $g'(x) = 3x$ az $a = x$ helyettesítés alapján. Az ekvivalensségről részletesen a Dieter Wickmann Bayes statisztika 133-ik oldalán lehet olvasni. Ez alapján:

$$g_{ekv}(a, x) = \begin{cases} 4x - a - 3x = -1(-x + a), & \text{ha } x \leq a \\ 3a - 3x = -3(x - a), & \text{ha } x \geq a \end{cases}$$

Ebből az előjelek megfordításával az $s(a, x)$ veszteségfüggvény jön létre:

$$s(a, x) = \begin{cases} a - x, & \text{ha } x \leq a \\ 3(x - a), & \text{ha } x \geq a \end{cases}$$

Ez a veszteségfüggvény általános formában is felírható.

$$s(a, \theta) = \begin{cases} k_{\uparrow}(a - \theta), & \text{ha } \theta \leq a \\ k_{\downarrow}(\theta - a), & \text{ha } \theta \geq a \end{cases}$$

A $k_{\uparrow}$ a túlbecsülés arányossági faktora, a $k_{\downarrow}$ pedig az alábecsülés arányossági faktora. Ezek alapján szeretnénk meghatározni a εs minimumát.

$$\varepsilon s = \int_{\theta} s(a, \theta) \cdot f(\theta) d\theta$$

A Wickmann (1995) könyv 135 és 136-ik oldalain a részletes levezetés megtalálható. A levezetés után:

$$\varepsilon s = (k_{\uparrow} + k_{\downarrow}) a F(a) - k_{\downarrow} a - (k_{\uparrow} + k_{\downarrow}) \phi(a) + k_{\uparrow} \phi(\theta_u) - k_{\downarrow} \phi(\theta_o)$$

Ahol, ϕ θ valamilyen függvénye, ez számunkra most nem érdekes. A minimum εs a szerinti deriváltját 0-val egyenlővé téve adódik.

$$\frac{d}{da} \varepsilon s = (k_{\uparrow} + k_{\downarrow})(1 \cdot F(a) + a \cdot f(a)) - k_{\downarrow} - (k_{\uparrow} + k_{\downarrow}) \phi(a) = (k_{\uparrow} + k_{\downarrow}) F(a) - k_{\downarrow}$$

Ha ezt egyenlővé tesszük nullával, akkor:

$$F(a_*) = \frac{k_{\downarrow}}{k_{\uparrow} + k_{\downarrow}}$$

Az általunk kapott értékeket behelyettesítve:

$$F_N(a_* | \mu, \sigma) = \frac{3}{1 + 3} = 0.75$$

Innen a_* kiszámolható:

$$a_* = F_N^{-1}(0.75 | 1000, 435) = 1293$$

Így megkaptuk, hogy 1293 adag narancs vásárlása esetén lesz a kereskedő nyereségfüggvénye a legnagyobb.

4.3. Mennyi is az annyi?

Az előző alfejezetben megállapítottuk, hogy 1293 adag narancs vásárlása esetén lesz a kereskedő nyereségfüggvénye a legnagyobb. De mekkora is ez a nyereség? Mekkora a különbség akkor, ha a kereskedő a jelenleg általa meghatározott heti 1000 adagot rendeli?

4.3.1. A nyereség kiszámítása

Először is szeretnénk meghatározni, hogy mekkora lesz a kereskedő nyeresége, ha 1293 adag narancsot rendel. A nyereségfüggvényünk a következő 4.1. alapján:

$$\varepsilon g(a) = \int_{-\infty}^{a} (4x - a) f_N(x) dx + \int_{a}^{\infty} (3a) f_N(x) dx$$

Ezt az integrált kell kiszámolnunk $a = a_* = 1293$-ra.

$$\varepsilon g(a) = 4 \int_{-\infty}^{a} x f_N(x) dx - a \int_{-\infty}^{a} f_N(x) dx + 3a \int_{a}^{\infty} f_N(x) dx$$

A második és harmadik tag könnyen kiszámolható:

$$\int_{-\infty}^{a} f_N(x) dx = F_x(a)$$

$$\int_{a}^{\infty} f_N(x) dx = 1 - F_x(a)$$

Az első tagot pedig a következő helyettesítés segítségével számolhatjuk ki:

$$z = \frac{x - \mu}{\sigma}, x = \mu + \sigma z, dx = \sigma dz, z_a = \frac{a - \mu}{\sigma}, f_N(z) = \sigma f_N(x)$$

Valamint használjuk a $f_N(x) = f_x$ és $f_N(z) = f_z$ jelöléseket. Így:

$$\int_{-\infty}^{a} x f_N(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_a} (\mu + \sigma z) e^{-\frac{1}{2}z^2} dz$$

$$= \mu F_z(z_a) + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{z_a} z e^{-\frac{1}{2}z^2} dz$$

$$= \mu F_z(z_a) - \sigma f_z(z_a) = \mu F_x(a) - \sigma^2 f_x(a)$$

Így behelyettesítve $F_x(a_x) = \frac{3}{4}$-t, $\mu = 1000, \sigma = 435$-t, valamint $f_x(a_* = 1293) = 7.3 \cdot 10^{-4}$ a következőt kapjuk:

$$\varepsilon g(a_*) = \cdot(1000 \cdot \frac{3}{4} - 435^2 \cdot 7.3 \cdot 10^{-4}) - 1293 \cdot (\frac{3}{4}) + 3 \cdot 1293 \cdot \frac{1}{4} \approx 2447$$

Tehát a kereskedő heti nyeresége ebben az esetben 2447.

4.3.2. Átlagkereslet szerinti nyereség

Ha az $a = 1000$ értéket helyettesítjük be az előző alfejezetben kiszámoltakba, akkor a következőt kapjuk:

$$F_x(1000) = 0.5, \mu = 1000, \sigma = 435, f_x(1000) = 9.17 \cdot 10^{-4}$$

$$\varepsilon g(a = 1000) = 4 \cdot (1000 \cdot \frac{1}{2} - 435^2 \cdot 9.17 \cdot 10^{-4}) - 1000 \cdot (\frac{1}{2}) + 3 \cdot 1000 \cdot \frac{1}{2} \approx 2306$$

Ezek alapján a kereskedő $2447 - 2306 = 141$ egységgel keresne többet, mintha az átlagkereslet szerint választaná a heti rendelését. Ez az egység a valóságban valamilyen pénznemet jelent (euró, forint, stb.).

5. fejezet

A gombák

5.1. Alapprobléma

A Bayes statisztika egy másfajta alkalmazását láthatjuk a következő probléma során. Tegyük fel, hogy adott egy adatbázis, melyben gombák vannak felsorolva. Minden gombának több különböző paramétere van. Számunkra most a legfontosabb az lesz, hogy ehető-e a gomba, vagy sem. Az eddigi adatok alapján jóslást szeretnénk arra adni, hogy ha megnézünk egy új gombát, ami nincs benne az adatbázisban, akkor az ehető lesz-e vagy sem. A probléma Vancsó Ödön és Princz Péter Integrating elements of data science into highschoool teaching: Naive Bayes classification algorithm, and programming in Python című cikkük alapján merült fel.

5.2. Naive Bayes osztályzás

5.2.1. Függetlenség

A Naive Bayes osztályozó nevében a "Naive" kifejezés arra utal, hogy független attribútumaink vannak, és azokat használjuk az osztályozás során. Érdemes megemlíteni, hogy az osztályozó egészen jól működik akkor is, ha nem függetlenek az attribútumok. A Naive Bayes féle osztályzó az osztályra vonatkozó feltételes valószínűséget becsüli meg, feltéve, hogy az attribútumok feltételesen függetlenek.

$$P(X|Y = y) = \prod_{i=1}^{d} P(X_i|Y = y)$$

ahol $X = \{X_1, X_2, ..., X_d\}$. (Tan, Steinbach & Kumar, 2006)

5.2.2. Az osztályozásról

Adott egy adathalmaz attribútumokkal $(A_1, A_2, ..., A_n)$. A célunk az, hogy megjósoljuk a C osztályt, pontosabban, hogy melyik C-re a $P(C|A_1, A_2, ..., A_n)$ a legnagyobb. Ezt a Bayes tétel segítségével a következőképpen tudjuk megtenni:

$$P(C|A_1, A_2, ..., A_n) = \frac{P(A_1, A_2, ..., A_n|C)P(C)}{P(A_1, A_2, ..., A_n)}$$

$P(C|A_1, A_2, ..., A_n)$-t maximalizálása ekvivalens azzal, ha olyan C-t választunk, ami maximalizálja $P(A_1, A_2, ..., A_n|C)P(C)$-t.
$P(A_1, A_2, ..., A_n|C)$-t ki tudjuk számolni, mivel feltettük, hogy függetlenek az attribútumok. Így:

$$P(A_1, A_2, ..., A_n|C) = P(A_1|C_j)P(A_2|C_j)...P(A_n|C_j)$$

(Tan, Steinbach & Kumar, 2006)

5.3. A feladat

5.3.1. Az adathalmaz

Az adathalmaz, amit a jóslásunkhoz szeretnénk felhasználni, a kaggle.com-ról származik. Ezen az oldalon rengeteg adathalmaz található, melyeket adatbányászathoz szoktak használni. Jelen esetben az adathalmazunk 8124 sorból és 24 oszlopból áll. A sorok tartalmazzák az egyes gombákat, az oszlopok pedig a gombák tulajdonságait (más néven attribútumait) tartalmazzák. (Megjegyzés: Az ábrán nem az összes oszlop látható)

	class	cap-shape	cap-surface	cap-color	bruises	odor	ring-number	ring-type	spore-print-color	population	habitat
0	p	x	s	n	t	p	o	p	k	s	u
1	e	x	s	y	t	a	o	p	n	n	g
2	e	b	s	w	t	l	o	p	n	n	m
3	p	x	y	w	t	p	o	p	k	s	u
4	e	x	s	g	f	n	o	e	n	a	g
...	...	...	...	...	...	...	...	...	...	...	...
8119	e	k	s	n	f	n	o	p	b	c	l
8120	e	x	s	n	f	n	o	p	b	v	l
8121	e	f	s	n	f	n	o	p	b	c	l
8122	p	k	y	n	f	y	o	e	w	v	l
8123	e	x	s	n	f	n	o	p	o	c	l

Az első oszlopban láthatjuk a gombák osztályba sorolását. Ez két fajta lehet, ehető (e), vagy mérgező (p). A további oszlopok pedig a gomba további tulajdonságaira vonatkoznak, például a gomba kalapjának alakja, a kalap bőre, a kalap színe, a gomba illata, előfordulása, stb.
Azonban mivel az adataink szöveges karakterként vannak megadva, így először ezeket a karaktereket át kell konvertálnunk számmá, hogy könnyebben dolgozhassunk velük. (Megjegyzés: Az ábrán nem az összes oszlop látható)

	class	cap-shape	cap-surface	cap-color	bruises	ring-umber	ring-type	spore-print-color	population	habitat
0	1	5	2	4	1	1	4	2	3	5
1	0	5	2	9	1	1	4	3	2	1
2	0	0	2	8	1	1	4	3	2	3
3	1	5	3	8	1	1	4	2	3	5
4	0	5	2	3	0	1	0	3	0	1
...	...	...	...	...	...	...	...	...	...	...
8119	0	3	2	4	0	1	4	0	1	2
8120	0	5	2	4	0	1	4	0	4	2
8121	0	2	2	4	0	1	4	0	1	2
8122	1	3	3	4	0	1	0	7	4	2
8123	0	5	2	4	0	1	4	4	1	2

5.3.2. A módszer bemutatása

A módszert az adathalmazunk egy leegyszerűsített verzióján mutatom be. 6 oszlopot fogunk vizsgálni, és 9+1 sort. 9 gomba alapján próbáljuk meg megjósolni, hogy a tizedik vajon ehető-e vagy sem. A következő táblázat mutatja az adatainkat:

Cap-shape	Cap-surface	Cap-color	Bruises	Odor	Class
5	2	4	1	2	1
5	2	9	1	2	0
0	2	8	1	3	0
5	3	8	1	2	1
5	2	3	0	3	0
3	2	4	0	0	0
5	2	4	0	0	0
2	2	4	0	0	0
3	3	4	0	7	1
5	2	4	0	2	?

A kérdőjel értékét szeretnénk megtudni. A Class oszlopban a 0 érték jelöli, hogy egy gomba ehető és az 1 érték, hogy mérgező. Jelölje Eß azt, hogy a gomba ehető, és M, hogy a gomba mérgező. Ezenkívül jelölje T a tulajdonságokat. Ezek után ki szeretnénk számolni a $P(T|E)$ értéket, és a $P(T|M)$.

$$P(T|E) = \frac{3}{6} \cdot \frac{6}{6} \cdot \frac{3}{6} \cdot \frac{4}{6} \cdot \frac{1}{6} = 0.028$$

$$P(T|M) = \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} = 0.033$$

$P(T|E)$ esetében a nevező 6, mivel 6 darab ehető gombánk van. A $\frac{3}{6}$ számlálójában pedig azért 3 áll, mivel a 6 ehető gombából 3-nak egyezik a kalap formája az általunk jósolni kívánt gomba kalap formájával. A többi érték hasonlóan jön ki mindkét esetben. Ezt az értéket meg kell szorozni az előfordulások valószínűségével:

$$P(T|E) \cdot P(E) = 0.028 \cdot \frac{6}{9} = 0.019$$

$$P(T|M) \cdot P(M) = 0.033 \cdot \frac{3}{9} = 0.011$$

Mivel $0.019 > 0.011$, így az általunk keresett gombát ehetőnek ítéljük. Természetesen ez egy nagyon leegyszerűsített példa, a valódi eljárásnál figyelembe fogjuk venni az összes oszlopot és sort. Viszont ezt már csak számítógéppel tudjuk kiszámolni.

5.3.3. A módszer végrehajtása a teljes adathalmazon pythonban

Az adathalmazunkat két részre bontjuk, a tanító és a teszt adathalmazokra. A tanító adathalmaz az egész adathalmaz 70%-ából fog állni, amit véletlenszerűen választunk ki. A teszt adathalmazra pedig azért lesz szükségünk, hogy megnézzük milyen pontos lesz a modellünk. Ezt úgy tudjuk megállapítani, hogy mivel a teszt adathalmaz eredményeit tudjuk, így azokat össze tudjuk vetni a jósolt értékekkel. A GaussianNB modellt fogjuk használni. Ez a modell feltételezi, hogy a mintánk normális eloszlású, valamint végrehajtja az előzőekben bemutatotthoz hasonlóan a számolást az adathalmazunkhoz. Először a tanító részt adjuk oda a modellnek, majd pedig a teszt részt, és végül összevetjük a jósolt értékeket a valós értékekkel.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.3,random_state=109)
```

```
>>> from sklearn.naive_bayes import GaussianNB
a = np.ravel(y_train)
gnb = GaussianNB()
gnb.fit(X_train, a)
y_pred = gnb.predict(X_test)
```

A teszt adathalmazon való futtatás után megkaptuk a jósolt értékeket. Most vessük össze a tényleges értékekkel ezeket:

```
from sklearn import metrics

print("Pontosság:",metrics.accuracy_score(y_test, y_pred))
```

```
Pontosság: 0.9196062346185397
```

Láthatjuk, hogy a modell 91,96%-ban jósolta meg jól, hogy a gomba ehető-e vagy sem.

6. fejezet

Zárszó

6.1. Végül egy izgalmas kérdés

Utoljára nézzünk még egy feladatot. Adott egy táblázat, melyben az autópályán egy 15 perces időintervallum alatt elhaladó járművek száma látható hat egymást követő napon, két héten keresztül.

	Hétfő	Kedd	Szerda	Csütörtök	Péntek	Szombat
Első hét	50	65	52	63	84	102
Második hét	56	49	60	45	112	90

Feltételezhetjük, hogy a járművek száma Poisson eloszlású. Így a hétfő és csütörtök között átlagosan elhaladó járművek száma legyen λ és a pénteken és szombaton elhaladó átlagos járművek száma pedig μ. Feltételezhetjük, hogy az egyes napok függetlenek. Adjunk meg olyan intervallumokat, amelyekbe a μ, illetve a λ paraméterek 95%-os eséllyel fognak beleesni. (Migon és Gamerman, 1999, p. 193, 12. feladat)

6.2. A feladat egy lehetséges megoldása

Először számoljuk ki ezt az intervallumot a λ paraméterre, azaz a hétfő és csütörtök közötti időszakra. Most csak az első hét adatait fogjuk figyelembe venni és azok alapján fogunk számolni. Írjuk fel a Poisson eloszlást és helyettesítsünk be az egyes napok értékeit:

$$P(x = 50) = \frac{\lambda^{50}}{50!} \cdot e^{-\lambda}, P(x = 65) = \frac{\lambda^{65}}{65!} \cdot e^{-\lambda}$$
$$P(x = 52) = \frac{\lambda^{52}}{52!} \cdot e^{-\lambda}, P(x = 63) = \frac{\lambda^{63}}{63!} \cdot e^{-\lambda}$$

Mivel az egyes mintavételek függetlenek, így az hét első négy egymás utáni napja alapján ezek a számok adódnak:

$$\frac{\lambda^{50}}{50!} \cdot e^{-\lambda} \cdot \frac{\lambda^{65}}{65!} \cdot e^{-\lambda} \cdot \frac{\lambda^{52}}{52!} \cdot e^{-\lambda} \cdot \frac{\lambda^{63}}{63!} \cdot e^{-\lambda} = \frac{\lambda^{230}}{50! \cdot 65! \cdot 52! \cdot 63!} \cdot e^{-4\lambda}$$

Ezt formálisan leírva:

$$P(x_1 = 50, x_2 = 65, x_3 = 52, x_4 = 63|\lambda = y) = \frac{y^{230}}{50! \cdot 65! \cdot 52! \cdot 63!} \cdot e^{-4y}$$

Viszont mi a következő valószínűséget keressük:

$$P(\lambda = y|x_1 = 50, x_2 = 65, x_3 = 52, x_4 = 63) = ?$$

Ezt a ki tudjuk számolni a Bayes-tétel segítségével:

$$P(\lambda = y|x_1 = 50, x_2 = 65, x_3 = 52, x_4 = 63) =$$

$$= \frac{P(x_1 = 50, x_2 = 65, x_3 = 52, x_4 = 63|\lambda = y) \cdot P(\lambda = y)}{\int_0^N P(x_1 = 50, x_2 = 65, x_3 = 52, x_4 = 63|\lambda = y)dy \cdot P(\lambda = y)}$$

Jelöljük λ sűrűségfüggvényét f-fel. Ekkor $P(\lambda = y) = f(y)$. Mivel f egyenletes eloszlású, így $f(y) = \frac{1}{N}$. Helyettesítsük be a kiszámolt értékeket:

$$P(\lambda = y|x_1 = 50, x_2 = 65, x_3 = 52, x_4 = 63) =$$

$$= \frac{\frac{y^{230}}{50! \cdot 65! \cdot 52! \cdot 63!} \cdot e^{-4y} \cdot \frac{1}{N}}{\int_0^N \frac{y^{230}}{50! \cdot 65! \cdot 52! \cdot 63!} \cdot e^{-4y} \cdot \frac{1}{N}}$$

Az egyszerűsítések után a következő törtet kapjuk:

$$\frac{y^{230} \cdot e^{-4y}}{\int_0^N y^{230} \cdot e^{-4y}}$$

Azt az $[a, b]$ intervallumot keressük, amibe az általunk keresett λ, 0.95 eséllyel beleesik. Ezt a következő egyenlőtlenség segítségével tudjuk kiszámolni:

$$\frac{\int_a^b y^{230} \cdot e^{-4y} dy}{\int_0^N y^{230} \cdot e^{-4y} dy} \geq 0.95$$

Számoljuk ki először a nevezőt, parciális integrálással:

$$\int_0^N y^{230} \cdot e^{-4y} dy = N^{230} \cdot \frac{e^{-4N}}{-4} - 0 \cdot \frac{e^{-4 \cdot 0}}{-4} - \int_0^N 230 \cdot y^{229} \cdot \frac{e^{-4y}}{-4} dy =$$

$$= N^{230} \cdot \frac{e^{-4N}}{-4} - \frac{230}{-4} \cdot \int_0^N y^{229} \cdot e^{-4y} dy$$

Láthatjuk, hogy az y fokát eggyel csökkenteni tudtuk. Így a parciális integrálást még 229 alkalommal el kell végeznünk, hogy megtudhassuk N függvényében a nevező értéket. Így:

$$\int_0^N y^{230} \cdot e^{-4y} dy = (N^{230} \cdot \frac{e^{-4N}}{-4} + \frac{230}{4} \cdot (N^{229} \cdot \frac{e^{-4N}}{-4} + \frac{230 \cdot 229}{4^2} (N^{228} \cdot \frac{e^{-4N}}{-4} + ...$$

$$.... + \frac{230 \cdot 229 \cdot ... \cdot 2}{4^{229}} (N \cdot \frac{e^{-4N}}{-4} - \int_0^N 1 \cdot \frac{e^{-4y}}{-4} dy)$$

Az utolsó integrálást is végezzük el:

$$\int_0^N e^{-4y} dy = (\frac{e^{-4N}}{-4} - \frac{1}{-4})$$

Így a kifejezést zárt alakra tudjuk hozni:

$$\int_0^N y^{230} \cdot e^{-4y} dy = N^{230} \cdot \frac{e^{-4N}}{-4} + \sum_{i=0}^{228} \frac{\prod_{j=0}^{i}(230-j)}{4^{i+1}} (N^{228-i} \cdot \frac{e^{-4}}{-4}) + (\frac{e^{-4N}}{-4} - \frac{1}{-4})$$

Egy N-től függő kifejezést kapunk így végeredményül a nevezőnek. Azonban ezt kiszámolni már számítógéppel lenne érdemes, egy analízises programcsomaggal. Azonban mivel ez egy N-től függő kifejezés, így meg kell becsülnünk azt a maximális értéket, amit N felvehet, és erre a becsült értékre, már a számítógép ki tudja számolni a kapott kifejezést. A tört számlálóját hasonlóképpen ezek alapján ki lehet számolni. Az pedig a-tól és b-től is fog majd függeni,és olyan a, b párt keresünk a definíció szerint, amelyre $b - a$ minimális és teljesül az egyenlőtlenség. Ezt approximációval lehet a nevezőhöz hasonlóan meghatározni számítógép segítségével.

6.3. Megjegyzés a használt értékekhez

Az eddigi számolásaink során csak az első hét értékeit vettük figyelembe. Érdemes lenne kiszámolni a másik hét szerinti intervallumot is, valamint a péntek-szombat értékei által meghatározott intervallumokat is. Utóbbiból a kérdésben szereplő mű értékére adódik majd egy intervallum. Más becsléssel is dolgozhatunk, ahol a megadott számoknak csak az átlagát használjuk (így ebben kevesebb információ kerül felhasználásra, de abból is lehet becsülni), de erre most idő és terjedelmi okból sem térek már ki.

6.4. Végszó

A dolgozat során belepillanthattunk, a fokozatosan egyre nehezebb feladatok kapcsán, hogy milyen is a Bayes-statisztika. Az alapoktól kezdve, ahol még kézi számításokkal is boldogultunk eljutottunk a nehezebb feladatokig is. Ezáltal láthatjuk a Bayes-statisztika összetettségét és mélységét. A gombás, illetve az imént látott feladat alapján láthatjuk, hogy a Bayes-statisztikában is előfordulhatnak nehéz számítások. Ezek elvégzésére számítógépes programokat kell használnunk.

Összességében megismerhettem és elmélyedhettem a Bayes-statisztika világában. Az ezen témabeli alapismereteimet a Bayes-tétel folytonos esetével, a Bayes-elvvel, különféle nyereségfüggvénybeli becslésekkel, osztályozó algoritmussal és sok más érdekes tapasztalattal bővíthettem. Remélhetőleg a jövőben sikeresen kamatoztathatom majd az így megszerzett tudásomat.

Irodalomjegyzék

- Bilecz R. (2020). *COVID-19 tudnivalók*, URL: https://www.mediteam.hu/covid.html (2020.12.03.)
- Centers for Disease Control and Prevention (2020). *Using Antigane Test*, URL: https://www.cdc.gov/coronavirus/2019-ncov/lab/resources/antigen-tests-guidelines.html (2020.12.03.)
- Dimenzió-Med Egészségközpont (2020). *COVID-19 tesztek*, URL: https://www.dimenzio.net/hirek/covid-19teszt/ (2020.12.03.)
- Fidy J., & Makara G. (2005).*Biostatisztika*, InforMed 2002 Kft.
- Havers, F., P., Reed, C., Lim, T., Montgomery, J., M., Klena, J., D., Hall, A., J.,... Thornburg, N., J. (2020). Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23-May 12, 2020, URL: https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2768834?guestAccessKey=7a5c32e6-3c27-41b3-b46c-43c4a38bbe00&utm_source=For_The_Media&utm_medium=referral&utm_campaign=ftm_links&utm_content=tfl&utm_term=072120 (2020.12.03.)
- Központi Statisztikai Hivatal (2020). *Népesség, népmozgalom(1941-)*, URL: http://www.ksh.hu/docs/hun/xstadat/xstadat_eves/i_wnt001c.html (2020.12.03.)
- Magyarország Kormánya (2020). *Koronavírus*, URL: https://koronavirus.gov.hu (2020.12.03.)
- Migon H., S., és Gamerman D. (1999). *Statistical Inference An Integrated Approach*, London: Arnold publisher
- Peabody R. (2019). *Sensitivity and specificity of HIV tests*, URL: https://www.aidsmap.com/about-hiv/sensitivity-and-specificity-hiv-tests (2020.12.22.)
- Synlab (2020). *Minden amit tudni kell a COVID-19 tesztekről*, URL: https://www.synlab.hu/storage/DyZJKMOirPOGoWjJLipbqE9xREZOSkGTG7OPR2Nv.pdf (2020.12.03.)
- Tan, P., Steinbach, M., & Kumar V. (2006). *Introduction to data mining*, Pearson Education, Inc.
- Vancsó Ö. (2005), *Klasszikus és Bayes-szazisztika a matematika didaktikában*, Debrecen

- Vancsó Ö., & Princz P. (2020). *Integrating elements of data science into highschool teaching: Naive Bayes classification algorithm and programming in Python*, Debrecen
- Wickmann D. (1999). *Bayes-statisztika*, Budapest:ELTE Eötvös Kiadó