Szakdolgozat

Kiss Zsófia
2012

Pont és intervallumbecslés bemutatása a Bayes- és a Klasszikus statisztika eszközeivel

Szakdolgozat

Kiss Zsófia
Matematika elemző szakirány
Témavezető:
Dr. Vancsó Ödön, egyetemi adjunktus
Matematika Tanítási és Módszertani Központ
Eötvös Lóránd Tudományegyetem, Természettudományi Kar

Budapest
2012

Tartalomjegyzék

1. fejezet

Bevezetés

1.1. Előszó

A Bayes statisztikával és alkalmazási területeivel először akkor találkoztam, amikor szabadon választható tárgyként fölvettem a Bayes-statisztika I. című tárgyat Dr. Vancsó Ödönnél. Már elsőre megfogott, hogy a saját szubjektív előítéleteimet is fölhasználhatom a becslések során. Sokkal életszerűbbnek és használhatóbbak tűnt számomra, mint a klasszikus statisztika. Nagyon furcsának találtam, hogy egészen eddig még csak említés szintjén sem találkoztam a Bayes-statisztikával, csak magával a Bayes- tétellel.

Később saját költségemre a Bayes-statisztika II. című tárgyat is fölvettem. Addigra már nagyon érdekelt, hogy mivel bővíthetem hiányos tudásomat. Illetve tanárom előadásmódja is magával ragadt. Nagyjából ekkor fogalmazódott meg bennem, hogy ebbe a témakörbe kívánom kiválasztani szakdolgozatom témáját. Szerencsére támogatást nyert elképzelésem.

Ezúton is szeretnék köszönetet mondani Dr. Vancsó Ödönnek, szakdolgozati konzulensemnek, aki időt és fáradtságot nem sajnálva aktívan részt vett szakdolgozatom létrejöttében. Segítsége nélkül nem sikerült volna!
Köszönöm!

1.2. A statisztika rövid történeti háttere

A statisztika mindig is lényeges része volt az emberek életének. Már az ókorban is jelentős szerepet kapott. A bibliában is megtalálhatjuk a statisztika kezdeti változatát: Máriának és Józsefnek Jeruzsálembe kellett utaznia, Mária szülővárosába. A születési hely szerint vezették a népességnyilvántartást, amely alapján előre meg tudták becsülni mekkora összeg fog befolyni az állam "kasszájába" a kivetett adókból. Így tervezhetővé vált a költségvetés.

A statisztika az olasz statiszta illetve a latin status (állapot, állam) szóból eredeztethető. Gottfried Achenwall 1749-ben elsőként alkalmazta a német állam tevékenységéből keletkezett adatok elemzésére. A szó a 19. századra új jelentést kapott. A matematikai statisztika a matematikai ismeretek kibővülésével, a valószínűségszámítás fejlődésével, a leiró statisztika és a különböző mintavételi technikák, adatgyűjtési módszerek segítségével fejlődött a mai szintjére.

A mai klasszikus felfogás az 1920-as, 30-as években alakult ki. Ekkor jött létre a fogalmi háttér. A mai klasszikus felfogásban két különböző gondolatkör fonódott össze. Az egyik Ronald A. Fisher nevéhez fűződik, aki angol természettudósként nagyon nagy hatással volt a statisztika történelmére. Fisher nem volt megelégedve a Bayes-féle szubjektív -priori valószínűségekkel, ezért szerette volna valamilyen módon eltűntetni az elméletből. Mivel a szubjektív valószínűségek használata ellentétben állt az objektív, személytelen felfogással. Élete során egyre nyitottabbá vált a Bayes-statisztikára.

A másik gondolatkör Jerzy Neyman és Egon Pearson nevéhez köthető. Ez a két szerző átdolgozta Fisher elméletét, amely éles ellentétben állt a 20-as, 30-as években szintén fejlődésnek indult Bayes-statisztikával, illetve Fisher-féle felfogással is (később a szignifikancia teszt kritikájánál részletesebben be fogjuk mutatni a Fisher - Neyman és Pearson ellentétet).

A véletlen mintavételt Fisher vezette be. A klasszikus felfogás fő gondolatmenete az, hogy meghatároz egy tartományt, amelybe adott hipotézis mellett a kísérlet eredményének egy előre meghatározott valószínűséggel esnie kell.

A klasszikus valószínűség-számítás jól ismert Bayes-tételén alapuló, Bayes-statisztikai megközelítés az utóbbi 15-20 évben a számítástechnika gyors fejlődésével kapott nagy lendületet. A modern Bayes-statisztika Bruno de Finetti, Jimmy Savage, Dennis Lindley és társaik munkássága során indult virágzásnak, ami az 50-es, 60-as évekre tehető. Mára már a klasszikus statisztika konkurens elméletévé vált. Az elmélet alapja egy jóval tágabb valószínűség fogalom, amit nem csak szubjektív valószínűség esetén alkalmazhatunk.

A Bayes-statisztika teljesen új szemlélete nagy vitát váltott ki a klasszikus statisztika követői és az új szemlélet pártolói között. Elsősorban a külső információk felhasználásán és a priori-eloszlás szubjektívitásán volt a vita hangsúlya, nem a formula helyességén, mert az kétségtelen. Sok esetben a kiindulási modell megválasztásán van a konfliktus alapforrása. Ami gyakran mindenféle alap, ellenőrzés nélkül történik. Emiatt sokan bizalmatlanok a kapott eredménnyel kapcsolatban. Ma már a kutatások és gyakorlati dolgok terén is egyre nagyobb népszerűségnek örvend a Bayes-i szemléletmód.

Ebben a szakdolgozatban a statisztikai becsléselmélet pont és intervallumbecslését fogom bemutatni a kétféle megközelítési módszerrel. Így szemléletesebbé válik majd a hasonlóság és a különbség a két módszer között.

1.3. Általános fogalmak bevezetése

Ahhoz, hogy a kétféle módszert bemutathassam, szükség van néhány alapfogalom bevezetésére, amit használni fogok.

Eseménynek nevezzük egy időhöz és helyhez kötött történést, mely vagy bekövetkezik vagy nem. A természeti törvények nem számítanak eseménynek, mivel általános állítások és nem kötöttek helyhez vagy időhöz.

Az **eseményvalószínűség** megadja, hogy egy adott fizikai rendszer az egyes állapotait mekkora valószínűséggel veszi föl. Azaz a vizsgálandó események számát osztjuk az összes lehetséges kimenetellel.

Kísérlet alatt egy esemény megfigyelését értjük. Véletlen kísérlet alatt olyan megfigyelést értünk, aminek kimenetele előre nem jósolható meg.

Egy véletlen kísérlet elemi tulajdonsága, hogy megismételhető, ideális esetben végtelen sokszor. Egy kísérletnek két kimenete lehet: vagy bekövetkezik, vagy nem. Ha egy kisérlet n-szeri ismétlésével k-szor bekövetkezik az esemény , akkor az esemény **relatív gyakorisága** $r_n = \frac{k}{n}$ az n kísérletben. Mivel az esemény bekövetkezése csak $k = 1, 2...n$ esetén lehetséges, így $0 \leq \frac{k}{n} \leq 1$ egyenlőtlenség teljesül.

A kísérletszám (n) folyamatos növelésével a relatív gyakoriság egy bizonyos p szám körül fog ingadozni. Az n növelésével ez az ingadozás egyre csökkenő mértékű, viszonylagos stabilitás vehető észre.(Nagy számok törvénye) Azaz az A eseményhez hozzárendelhetünk egy valós számot, amely körül a relatív gyakoriság ingadozás viszonylagos stabilitása észlelhető. Ezt az A esemény valószínűségének nevezzük és $P(A)$-val jelöljük. Ekkor definiálhatunk egy P függvényt , amelynek $P(A) = p$ helyettesítési értéke adja az esemény valószínűségét. A valószínűség egy esemény bekövetkezésének gyakoriságára jellemző számérték. A relatív gyakoriság egyenlőtlensége miatt itt is érvényes, hogy $0 \leq P(A) \leq 1$.

A valószínűség, a véletlen kísérlet objektív sajátosságaként általában ismeretlen, mert nem vagyunk képesek végtelen kísérletet elvégezni, hisz világunk véges.

Az r_n gyakoriság emiatt mindig csak egy közelítő becslése lehet p-nek. Az a kérdés, hogy "mi a jó becslése p-nek" a valószínűségszámítás jellemző vizsgálati tárgya.

Objektívnek nevezünk valamit, ha úgy véljük tőlünk függetlenül létezik. Ebben az értelemben az objektív szó fogalma a külső világ sajátosságaként jelenik meg. Egy véletlen esemény bekövetkezésének valószínűsége szintén tőlünk független, a kísérlet egyéni tulajdonságaként tekinthető. Egy véletlen rendszer sajátja. Ezt nevezzük **objektív valószínűségnek**.

Állapotnak nevezünk egy olyan valamit amiről szeretnénk megtudni, hogy fönnáll-e vagy sem. Az esetek túlnyomó részében nem lehet biztosan állítani, hogy adott állapot fennáll-e vagy sem, ennek ellenére valamennyire biztosak vagyunk a fennállásában. Bármi lehet állapot. Bizonyos helyzetben bizonytalanság uralkodhat benne.

Egy ember adott állapot fönnállására vonatkozó bizonyossági fokát **szubjektív valószínűségnek** nevezzük.

A valószínűségszámítást axiómatikusan szokták felépíteni. A legismertebb talán a **Kolmogorov-féle axiómarendszer**. Itt formálisan vannak rögzítve a fogalmak, melyek eleget tesznek az axiómarendszernek.

Az adott esemény esetén elképzelhető állapotok halmazát $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ **állapottérnek** nevezzük. $P(\theta_j)$ az a valószínűség, amivel egy ember hiszi, hogy θ_j állapot áll fenn. Az összes lehetséges j értékre adott $P(\theta_j)$ érékek egy **valószínűségeloszlást** határoznak meg. Ez egy adott személy θ állapottérre vontatkozó becslését adja.

A $P_1(\theta_1)$ minta előtti valószínűségeloszlást, mely a kísérlet előtti értékítéletünket tartalmazza **priori eloszlásnak** nevezzük. A priori eloszlás mindig szubjektív.

A kísérlet eredményével befolyásolt $P_2(\theta_2)$ priori eloszlást, **posteriori eloszlásnak** hívjuk. A posteriori eloszlás kiszámítása a Bayes-tétel segítségével történik. Bayes-tételét a későbbiek folyamán részletesebben fogom tárgyalni.

Egy véletlen számadat jellemzéséhez tudnunk kell, hogy milyen értékek jöhetnek számításba, és mekkora valószínűséggel. Ezeket a véletlentől függő mennyiségeket **valószínűségi változónak** nevezzük.

A megszámlálható értékkészletű valószínűségi változókat **diszkrét valószínűségi változónak** hívjuk.

Legyen $F(x) = P(\eta < x)$ az η valószínűségi változó **eloszlásfüggvénye**. Ha η és ζ valószínűségi változók 1 valószínűséggel egyenlők, akkor eloszlásfüggvényeik azonosak. Bármely valószínűségi változó eloszlásfüggvénye monoton nem csökkenő, balról folytonos függvény, melynek határértéke $-\infty$-ben 0-a, ∞-ben pedig 1.

Egy η valószínűségi változó **sűrűségfüggvényének** nevezzük az $f(x)$ függvényt, ha ezzel a η valószínűség változó $F(x)$ eloszlásfüggvényét a következőképpen határozhatjuk meg:

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Ha η-nak létezik sűrűségfüggvénye , akkor $F(x)$ folytonos. Ilyenkor η-t **folytonos valószínűségi változónak** nevezzük.

2. fejezet

Statisztikai becslések elmélete

Mielőtt a pont és intervallumbecslést bemutatnám, fontosnak érzem megismertetni az olvasót a statisztikai becslések elméletével. Hisz a pont és intervallumbecslés is a statisztikai becslések közé tartozik.

2.1. Statisztikai becslés

A statisztikai becslésnél[1] a valószínűségeloszlások ismeretlen paramétereit szeretnénk meghatározni az ismert adatokból, úgy hogy adott módszerrel közelítünk. Ha egy fizikai mennyiséget szeretnénk megkapni hibát is tartalmazó mért adatokból, akkor a valószínűségeloszlás valamelyik sajátos paraméterének becslésére van szükség. Tudjuk, hogy egy X valószínűségi változó vagy valamilyen ismert pl. normális eloszlású vagy normális eloszlással közelíthető.
A sűrűségfüggvénye ekkor:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

ahol σ, m ismeretlen paraméterek, amit becsülni kívánunk a statisztikai

[1] Obadovics J. Gyula :Valószínűségszámítás és Matematikai statisztika című könyve szolgált a statisztikai becslés című fejezet forrásaként

mintából. A függvény függ tehát a paraméterek értékétől, azaz:

$$f(x; m; \sigma)$$

Ha valamelyik ismeretlen paramétert egyetlen számmal becsüljük akkor pontbecslésről, ha pedig intervallummal, amely meghatározott valószínűséggel tartalmazza a paramétert, akkor intervallumbecslésről beszélünk.

Mindkét esetben fontos, hogy jó becslést kapjunk. A jó becsléshez pedig szükségünk van bizonyos feltételek meglétére.

2.2. A jó becslés 4 feltétele

A pont és intervallumbecslésnél is nélkülözhetetlen, hogy megfelelő minőségű legyen a kapott információ. Ezalatt azt értem, hogy a kapott érték a mért adatok várható értéke körül kell, hogy legyen, lehetőleg kicsi szórással. Ha meg is egyezik a várható értékkel, akkor torzítatlan becslést kaptunk. Azonban nem minden paraméter esetén lehetséges torzítatlan becslést kapni. Néha elég nekünk, ha csak határértékben tart hozzá. Ekkor aszimptotikusan torzítatlan becslésről beszélünk.

Ha nem a várható érték körüli eredményt kaptunk valószínűsíthető, hogy valami számolási hibát követtünk el, vagy valamelyik mért adat túlságosan kiugró értékkel rendelkezik a többséghez képest. Így olyan értéket kaphat a becslésünk, mely nem valósághű.

Lényeges, hogy a becslésünk a lehető legjobb legyen a többi közül. Azaz olyat választunk, amelynek a szórása kisebb minden lehetséges érték mellett a többi torzítatlan becsléshez képest (hatásosság).

Fontos kívánalom, hogy a mintaelemszám növelésével a becslés értéke sztochasztikusan a valódi paraméterhez konvergáljon (konzisztencia).

Továbbá szeretnénk, ha a becslésünk minden információt tartalmazna a becsült értékre vonatkozóan (elégségesség). A gyakorlatban néha persze megelégszünk azzal, ha a főbb információkat tartalmazza.

Lentebb ezeket a fogalmakat fogom definiálni.

1. **Torzítatlanság**:

   egy $a_i = a_i(X_1, X_2, \ldots, X_n)$ statisztikát egy b paraméter torzítatlan becslése, ha a várható érték megegyezik b-vel. Ekkor:

   $$E(a_i(X_1, X_2, \ldots, X_n) = b$$

   Azaz az ingadozó értékek átlaga $a_i = a_i(X_1, X_2, \ldots, X_n)$ a paraméter valódi értéke körül kell, hogy szórodjék. Azonban léteznek olyan paraméterek, melyeknek nem létezik torzítatlan becslése.

Ekkor olyan becslést keresünk, melyek elég nagy n esetén már elég kicsi torzítással rendelkeznek. Az $a_i = a_i(X_1, X_2, \ldots, X_n)$ statisztika aszimptotikusan torzítatlan becslése b paraméternek, ha minden lehetséges értékre :

$$\lim_{n\to\infty} E(a_i(X_1, X_2, \ldots, X_n) = b$$

Időnként előfordul, hogy az eloszlásfügvény több lehetséges paramétertől függ, és mi ezek egyikét vagy ezeknek a függvényét szeretnénk becsülni. A torzítatlanság definiciója persze ilyenkor is változatlan.

2. **Hatásosság (efficiencia)**:

   ha egy $a_i = a_i(X_1, X_2, \ldots, X_n)$ statisztika ingadozása megfelelően kicsiny, elhanyagolható, akkor a becslést hatásosnak mondjuk. A leghatékonyabb becslés az, melynek a szórásnégyzete más torzítatlan becslés szorásnégyzeténél kisebb. Megmutatható, hogy csak egyetlenegy ilyen létezik. Ha nincs ilyen, akkor az összes torzítatlan becslés szórásnégyzetének biztosan van legnagyobb alsó korlátja (infimuma). A becslés hatásfokán (H_f) a kérdéses becslés szórásnégyzetével osztjuk az összes lehetséges szórásbecslés alsó korlátját, ahol a hányados (0,1] tartományban mozog.

   $$H_f = \frac{inf D^2(a_i')}{D^2(a_i)}$$

   1-el csak akkor egyezik meg, ha a_i a legjobb torzítatlan becslése b-nek.

3. **Konzisztencia (megegyezés)**:

   egy a_i statisztika sorozat valamely b paraméter konzisztens becslése, ha elég nagy n esetén az a_i sorozat elegendően nagy valószínűséggel jól közelíti a paraméter értékét. Azaz:

   $$\forall \varepsilon, \delta > 0 \quad \exists N \quad P(|a_i(X_1, X_2, \ldots, X_n) - b| \geq \varepsilon) \leq \delta \quad ha \qquad n > N$$

   Ha a becslés torzítatlan és szórásnégyzete n növekedésével 0-hoz tart, akkor b-nek erősen konzisztens becslése az a_i.

A konzisztencia az erősen konzisztensnél gyengébb követelmény. Belátható, hogy ha egy statisztika erősen konzisztens, akkor konzisztens is.

A konzisztens becslés nem feltétlenül torzítatlan, de a torzítás mértéke n növekedésével egyre csökkenő és 0-hoz tart. Ha lehetőségünk van elegendően nagy mintavételre, elegendő lehet egy olyan konzisztens becslés is ami nem torzítatlan. A torzítatlan becsléseknek a kis mintavételeknél van nagyon nagy jelentőségük.

4. **Elégségesség**:

   ha az a_i eloszlása minden információt tartalmaz a kérdéses paraméterre, akkor $a_i = a_i(X_1, X_2, \ldots, X_n)$ függvényt elégséges statisztikának hívjuk. A matematikai statiszikában az elégséges statisztikát használják általában. Azonban előfordulhat, hogy a könnyebb számolás kedvéért lemondunk a maximális információról és inkább egyszerűbb alakú statisztikát alkalmazunk.
   Elégséges statisztika önmagában nem létezik, mindig valamilyen paraméternek a függvényében lehet csak elégséges egy statisztika. Elégséges statisztika lehet az eredeti adatok használata is, elvégre minden lehetséges információt tartalmaznak, de mi mindig a lehető legszűkebbet keressük.

Megjegyzések:

- Minden eloszlásnál a mintaelemek mintaközepe torzítatlan becslést ad a várható értékekre. A korrigált tapasztalati szórásnégyzet torzítatlan becslése az elméleti szórásnégyzetnek.

- Normális, Poisson és exponenciális eloszlás várható értékére elégséges becslés a mintaközép

- Sűrűségfüggvény becslését sűrűséghisztogrammal végezzük

3. fejezet

Bayesianus becslés

A Bayes becslés alapvetően Bayes tételén alapul. Így elengedhetetlen annak ismerete.

3.1. Bayes-tétel

A valószínűségszámtás egyik lényeges eleme a feltételes valószínűség fogalma. Számszerűsített értéke egy jól meghatározott képlettel adható meg. A formula meghatározására különböző felírásmód is ismert. Az egyik ilyen a Bayes-tétele.

A Bayes-tétel megalkotása az angol methodista lelkész Thomas Bayes (1702-1761) nevéhez köthető. Bayes maga sosem használta a szubektív valószínűséget, a klasszikus felfogás híve volt. Sosem publikálta tételét. Halála után 1763-ban barátja Richard Price hozta nyilvánosságra a Philosophical Transaction of the Royal Society című munkában.

Fontos megemlítenünk, hogy lényeges eltérés van Thomas Bayes eredeti megoldása és a mai Bayes-statisztikai megoldás között. Bayes eleve egy olyan szituációt elemzett, ahol a helyzetből adódott a priori egyenletes-eloszlás. Ő nem gondolkozott szubjektív valószínűségekben. Az ő idejében a szerencsejátékokról, vagy más véletlen mechanizmusokról az egyenletes eloszlás objektív feltételezés volt.

Bayes elmélete csak az alábbi szituációban használható fel:
Egy véletlen eljárással kiválasztunk egy urnát a meglevő urnák közül, melyekben különböző arányú fekete és fehér golyó van. Visszatevéses mintavétellel kihúzunk belőle pár golyót. Arra vagyunk kiváncsiak, hogy melyik urnát is választottuk a meglévők közül. Ekkor valóban a kiválasztás véletlenszerű, tehát van értelme a priori eloszlásnak, amely pont a véletlen mechanizmusra irányul. Itt semmilyen szubjektív elem sincs a választásban.

Bayes maga csak ilyen helyzetekben használta tételét. Később a valószínűség tágabb értemlezésével kapta jelenlegi szerepét az adott tétel a Bayes-statisztikában.

A Bayes-tétel valójában egy képlet, amely képes tesz minket arra, hogy egy A esemény bekövetkezéséből le tudjuk vonni a konzekvenciát a lehetséges $B_1, B_2, \ldots, B_n$ események valószínűségére. Így megtudhatjuk, hogy A esemény bekövetkezésével milyen mértékben járul hozzá a hipotézisek cáfolásához vagy megerősítéséhez.

Ezért ismernünk kellene a priori (hipotézis észlelése előtti) valószínűségeket. A való életben ezek nagy részt nem ismertek, és hiba jöhet létre azzal, hogy magunk adunk értékeket nekik.

A Bayes-tétel megfelelő használata segíthet az előrejelzések esetén, ahol a feltevések hibáit szeretnék kiküszöbölni.

Bayes-tétele nem csak objektív, hanem szubjektív valószínűségek esetén is alkalmazható.

3.1.1. Bayes tételének bemutatása

A feladatunk, hogy megbecsüljünk egy adott esemény, vagy egy modell feltételes valószínűségét a meglévő információink és a megfigyelési adatok segítségével. Ha az előismereteinket fölhasználva becsülünk prediktív (előrejelző) becslésről beszélünk. Amikor a mért adatokat használjuk föl, akkor parametrikus becslést alkalmazunk. Mindkét esetben Bayes tételéből indulunk ki.

Bayes tétele: Az A esemény bekövetkezéséből szeretnénk B_i esemény valószínűségére következtetni. Ha a $B_1, B_2, \ldots, B_n$ események teljes eseményrendszert alkotnak (események egy sorozata, mely egymást páronként kizárja és összegük kiadja az egész eseménytért), és $P(B_i) > 0, i = 1, 2, \ldots, n$ és A egy tetszőleges pozitív valószínűségű esemény, akkor:

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)} \tag{3.1}$$

A tételben található $P(B_1), P(B_2), \ldots, P(B_n)$ kiindulási valószínűségeket priori valószínűségnek nevezzük. Ezek szubjektívek, és emiatt alakult ki a vita a klasszikus és a bayesianus hívek között. Az A esemény ismeretében csak azok a valószínűségek fognak változni, melyek A-tól függenek.

A $P(B_1|A), P(B_2|A), \ldots, P(B_n|A)$ feltételes valószínűségeket posteriori valószínűségnek hívjuk.

A Bayes tétel gyakorlati alkalmazása során adott, a kísérlet eredményétől nem függő $P(B_i)$ valószínűségekből indulunk ki. Majd a kísérlet során nyert információkat felhasználva meghatározzuk $P(B_i|A)$ valószínűségeket. Emiatt a $P(B_i)$ ismeretére már a kísérlet végrehajtása előtt szükségünk van, ezért tekintjük adottnak. A $P(B_i)$ valószínűségek megadására csak a kisérlet végrehajtása után van lehetőségünk.

A posteriori eloszlás a modell strukturájának paraméterezésére vagy magára a struktúrára is kiszámíthatjuk. Paraméterezés esetén a képlet formailag megegyezik a fenti Bayes képlettel.

Bayes- tétele véletlen változók esetén:[1] Vegyünk egy eseményrendszert, ahol X és Y véletlen változók.
(Bayes tételében:$A = (X = x); B = (Y = y)$).
Ekkor három eset lehetséges:

- X folytonos valószínűségi változó és Y diszkrét:

$$f_X(x|Y = y) = \frac{P(Y = y|X = x)f_X(x)}{P(Y = y)} \tag{3.2}$$

- X diszkrét és Y folytonos valószínűségi változó:

$$P(X = x|Y = y) = \frac{f_Y(y|X = x)P(X = x)}{f_Y(y)} \tag{3.3}$$

- X és Y is folytonos:

$$f_X(x|Y = y) = \frac{f_Y(y|X = x)f_X(x)}{f_Y(y)} \tag{3.4}$$

Bayes tétele folytonos esetre [2]:
Legyen ζ, η két valószínűségi változó, melyek együttes eloszlása abszolút folytonos a $h(x, y)$ sűrűségfüggvényel. Továbbá legyen :

$$f(x) = \int_{-\infty}^{\infty} h(x, y)dy \tag{3.5}$$

$$g(y) = \int_{-\infty}^{\infty} h(x, y)dx \tag{3.6}$$

$$f(x|y) = \frac{h(x, y)}{g(y)}, \text{ ha } g(y) > 0, \text{egyébként tetszőleges} \tag{3.7}$$

$$g(y|x) = \frac{h(x, y)}{f(x)}, \text{ ha } f(x) > 0, \text{egyébként tetszőleges} \tag{3.8}$$

$$\text{Ekkor:} f(x) = \int_{-\infty}^{\infty} f(x|y)g(y)dy \tag{3.9}$$

$$g(y) = \int_{-\infty}^{\infty} g(y|x)f(x)dx \tag{3.10}$$

[1] forrásul a [2]-es szakirodalom szolgált

[2] Rényi Alfréd: Valószínűségszámítás című könyve szolgált ennek a résznek a forrásaként

A (3.5),(3.6) képletekből megkaphatjuk:

$$g(y|x) = \frac{f(x|y)g(y)}{f(x)}, \text{ ha } f(x) > 0, \tag{3.11}$$

a (3.7) felhasználásával:

$$g(y|x) = \frac{f(x|y)g(y)}{\int_{-\infty}^{\infty} f(x|t)g(t)dt}. \tag{3.12}$$

A (3.10) képletre a Bayes-tétel általánosításaként tekinthetünk, folytonos eloszlások esetén.

3.2. Bayes - becslés

Pontbecslésnek nevezzük, ha egyetlen becsült értéket akarunk megadni. Intervallumbecslésnek, ha a becsült értékek egy tartományt vagy intervallumot határoznak meg. A pontbecslés során a posteriori eloszlás maximumát keressük. Intervallum-becslésnél olyan szűk intervallumot, amelybe előre meghatározott valószínűséggel esik a becsülendő valószínűség.

Sokféle pont és intervallum-becslés létezik. A becslés folyamatát befolyásolhatja, hogy milyen érdekek vezérelnek minket. Szeretnénk, ha várható értékben adná meg a paramétert, vagy mást kívánunk elérni. Folytonos, vagy diszkrét változókkal kívánunk számolni...

Minden lehetséges eljárási folyamatot nem tudunk ebben a szakdolgozatban bemutatni, hisz rengeteg variáció van. Úgy gondolom sokkal könnyebb egy módszert megérteni, ha gyakorlati példán kereszül látjuk a folyamatát. Így itt most egy folytonos esetre vonatkozó becslés megoldását fogjuk kiszámolni. A későbbiek folyamán a klasszikus becslésnél is ezt a példát fogjuk fölhasználni, hogy összehasonlíthatóak legyenek az adatok.

3.3. Gyakorlati példa

Meg kell említenem, hogy ez egy átdolgozott példa. A magyar lottóra vonatkozó számítások és az arra vonatkozó általánosítás megtalálható Dr. Vancsó Ödön: Klasszikus és Bayes-i statisztika a matematika didaktikában című doktori disszertációjában. Erre a feladatra vonatkozó számolásokat én végeztem.

Tegyük fel, hogy él New York-ban egy jóbarátunk. Vett egy lottót és megigéri nekünk, hogy ha nyer rajta megfelezi a nyereményét velünk. Mi semmit sem tudunk erről a lottóról csak a számokat látjuk a szelvényen. Szeretnénk megtudni hány számból húzzák ki, hogy meg tudjuk becsülni a nyerési esélyeket.

Az alábbi számokat látjuk: 2012.05.19-ei lottóeredmények

$$2, 17, 22, 30, 35, 49|45$$

Látható, hogy összesen 7 számot húznak ki és ebből az utolsó egy pótszám. Ekkor az alábbi lesz a matematikai modellünk: Legyen N a legnagyobb ismeretlen lottószámunk, és jelölje az $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ a kihúzott lottószámokat növekvő sorrendben. Jelöléseknek megfelelően látható, hogy N csak is nagyobb lehet a 7 kihúzott szám maximumával, vagy azzal megegyező. Sokféle közelítési módszer létezik, mely a legnagyobb kihúzott szám alapján próbál közelíteni. Ezek közül mi csak egyet fogunk bemutatni. Azért ezt a módszert fogjuk választani, mert várható érték hűvé tehető, tehát torzítatlan, konzisztens és effektív. A szórása minimális.

Direkt valószínűségszámítási megközelítést fogunk alkalmazni. Azaz ismerve a legnagyobb lottószám értékét, mekkora a valószínűsége annak, hogy a maximuma éppen x_7 lesz. Ekkor az összes lehetőség közül: $\binom{N}{7}$, meg kell számolnunk azokat a helyeket, ahol a maximum éppen x_7. Ekkor a keresett eseményvalószínűség az alábbi lesz: $\frac{\binom{x_7-1}{6}}{\binom{N}{7}}$. Itt már látható, hogy rögzített x_7 mellett N-ben akkor lesz maximális, ha a nevező minimális, ami $N = x_7$ esetén igaz. Mi viszont azt szeretnénk vizsgálni, hogy ha N ismeretlen, akkor ismert x_7-ből milyen következtetésekre juthatunk.

Először is szükségünk lesz egy priori állapot-valószínűség eloszlásra a lehetséges N értékekre. Mivel itt most semmilyen más infomációnk sincs, így egyenletes eloszlást feltételezünk egy lehetséges K maximális értékig. Ha erről se tudunk semmit, feltételezhetjük, hogy a végtelenhez tart. Bayes tételébe behelyettesítve az alábbit kapjuk:

$$P(N = Z|max = x_7) = \frac{P(max = x_7|N = Z)P(N = Z)}{\sum_{I=x_7}^{K} P(max = x_7|N = I)P(N = I)}$$

A priori egyenletes állapotvalószínűségek miatt $P(N = I)$ értékek azonosak. Mivel mind a számlálóban mind a nevezőben jelen vannak, így azokkal bátran leegyszerűsíthetünk.
Ekkor:

$$P(N = Z|max = x_7) = \frac{P(max = x_7|N = Z)}{\sum_{I=x_7}^{K} P(max = x_7|N = I)} =$$

$$= \frac{\frac{\binom{x_7-1}{6}}{\binom{Z}{7}}}{\sum_{I=x_7}^{K} \frac{\binom{x_7-1}{6}}{\binom{I}{7}}} = \frac{1}{\binom{Z}{7}} * \frac{1}{\sum_{I=x_7}^{K} \frac{1}{\binom{I}{7}}}$$

egyenlőséghez jutunk, amely már egy posteriori állapot-valószínűség eloszlás lesz a lehetséges $x_7, x_7 + 1, \ldots, K$ értékeken.

Látható, hogy a nevezőjében összeget tartalmazó szorzótényező konstans, mert nem függ a Z változótól. Csak egy normáló tényező, hogy eloszláshoz jussunk. Az első tényező Z növekedésével lecsökken, mivel a nevező növekszik, emiatt az eloszlás monoton csökkenő.

3.3.1. Pontbecslés

Mivel az előbb megállapítottuk, hogy monoton csökkenő a fenti megközelítésünk, ezért a legvalószínűbb Z érték éppen $Z = x_7 = 49$ lesz. Ez a poszteriori állapotvalószínűség-eloszlás maximuma. Későbbiek folyamán látni fogjuk, hogy klasszikus pontbecsléssel is erre az eredményre fogunk jutni. De a klasszikus és a bayes-i becslés által kapott értékek között egyenlőség csak akkor áll fent, ha speciális feltételek teljesülnek. Különben különbözni fognak az eredmények. Ezt az összehasonlításban részletesebben fogjuk elemezni.

3.3.2. Intervallumbecslés

A feladatunk, hogy meghatározzunk egy $0,95$ valószínűségű Bayes-tartományt. Ehhez módosítanunk kell a normáló tényezőt. Legyen:

$$\sum_{I=x_7}^{K} \frac{1}{\binom{I}{7}} = \frac{7}{6}\left(\frac{1}{\binom{x_7-1}{6}} - \frac{1}{\binom{K}{6}}\right) \tag{3.13}$$

A posteriori eloszlás ekkor:

$$P(N = Z|max = x_7) = \frac{1}{\binom{Z}{7}} * \frac{1}{\sum_{I=x_7}^{K} \frac{1}{\binom{I}{7}}} = \frac{1}{\binom{Z}{7}} * \frac{6}{7} * \frac{\binom{x_7-1}{6} * \binom{K}{6}}{\binom{K}{6} - \binom{x_7-1}{6}}$$

lesz, ahol Z változik az adottnak vett x_7 és K között. Mivel monoton csökken az eloszlás x_7-től kezdve, ezért a **legsűrűbb Bayes-tartomány**[3] egy$[x_7, T]$ tartomány lesz, ahol T meghatározása a

$$\sum_{Z=x_7}^{T} \frac{1}{\binom{Z}{7}} * \frac{6}{7} * \frac{\binom{x_7-1}{6} * \binom{K}{6}}{\binom{K}{6} - \binom{x_7-1}{6}} \geq 0,95$$

egyenlőtlenséget kielégítő legkisebb T érték.

Egyszerűsítsük tovább az egyenlőtlenségünket a könnyebb számolhatóság érdekében. Ejtsük ki K értékét anélkül, hogy kiszámolnánk. Nézzük meg mi történik, ha K tart a végtelenhez. Ekkor a harmadik szorzótényező számlálóját és nevezőjét is egyszerűsíthetjük K alatt a 6-tal, s alkalmazva, hogy K tart a végtelenhez megkapjuk, hogy határesetben a legkisebb T megoldás, amelyre:

$$\sum_{Z=x_7}^{T} \frac{1}{\binom{Z}{7}} * \frac{6}{7} * \binom{x_7 - 1}{6} \geq 0,95$$

[3] **Legsűrűbb Bayes-tartománynak** nevezzük azt a tartományt, amelyre két feltétel teljesül:

- Annak az esélye, hogy a véletlen szám ebbe a tartományba esik 0,95; 0,99, vagy még ennél is nagyobb, de 1-nél kisebb érték.
- Ez a tartomámy az előbbi feltételnek eleget tevők közül a lehető leszűkebb.

erre ismét fölhasználva a (3.11)-es képletet:

$$\frac{6}{7}*\binom{x_7-1}{6}*\sum_{Z=x_7}^{T}\frac{1}{\binom{Z}{7}}=\frac{6}{7}*\binom{x_7-1}{6}*\frac{7}{6}*\left(\frac{1}{\binom{x_7-1}{6}}-\frac{1}{\binom{T}{6}}\right)=1-\frac{\binom{x_7-1}{6}}{\binom{T}{6}}\geq 0,95$$

amiből már megkaphatjuk, hogy $0,05\binom{T}{6}\geq\binom{x_7-1}{6}$ egyenlőtlenséget kielégítő T érték lesz a keresett intervallum legfelső határa.

Ebből megkapjuk, hogy:

$$T(T-1)(T-2)(T-3)(T-4)(T-5)\geq 20(x_7-1)(x_7-2)(x_7-3)(x_7-4)(x_7-5)(x_7-6)$$

teljesüljön T-re.

Ekkor már csak be kell helyettesítenünk a lottósorsolás legmagasabb értékét az x_7 helyére, majd a jobb oldal 6-ik gyöke körüli T értéket próbálgatva (általában növelni kell az értéket) megkapjuk az intervallum felső határát.

Most térjünk vissza az eredeti példánkhoz és helyettesítsünk vissza az értékeket: 2012.05.19-ei lottóeredmények

$$2,17,22,30,35,49|45$$

a fenti képletbe behelyettesítve:

$$T(T-1)(T-2)(T-3)(T-4)(T-5)\geq 20(49-1)(49-2)(49-3)(49-4)(49-5)(49-6)=$$

$$=20*48*47*46*45*44*43=176709772800$$

Ennek vesszük a 6-ik gyökét, ami $74,9$ lesz. Fölfelé kerekítünk, mert 74 6-ik hatványa nem adja vissza az eredeti értéket. Tehát a becsült intervallumunk: $(49;75)$ lesz.

Nézzük meg, hogy 10 lottósorsolás maximális értéke közül mennyi esik majd a konfidencia intervallumunkba. A New York-i lottó 2002 óta létezik, és minden szerdán és szombaton sorsolnak 59 számból 7-et.

Minden évből választottam véletlenszerűen egy sorsolást 2011-ig:

időpontok	kihúzott számok
2002.05.25	15,10,16,28,34|49
2003.03.12	4,31,35,42,46,58|40
2004.12.29	19,20,41,42,44,54|23
2005.11.16	12,23,27,31,58,59|52
2006.01.04	10,15,20,25,38,51|43
2007.08.25	2,4,24,31,36,50|51
2008.09.13	7,14,25,27,39,41|53
2009.10.07	3,5,20,31,57,59|28
2010.07.03	4,7,19,34,41,56|39
2011.04.02	15,18,19,24,41,59|10

A legnagyobb számok: $49, 58, 54, 59, 51, 51, 53, 59, 56, 59$.A maximális érték 10-ből 10-szer benne volt a konfidencia intervallumunkba, ami nem is csoda, hisz csak akkor nem kerül bele az érték a becsült intervallumunkba, ha 49-nél kisebb a legmagasabb kihúzott szám értéke.

3.4. Bayes közelítés előnyei

- A paraméterbecslésnél, ha csak egyedül az adatokból következtetünk a paraméterekre (nem visszük bele az egyéni megérzéseinket), akkor a következtetések levonását a hipotetikus (feltételezett) viselkedés figyelmen kívül hagyásával végzi a Bayes-i közelítés.

- A paraméterezés bizonytalanságát eloszlással jellemezhetjük, ami egy többlépcsős tanulási folyamatnál, vagy tudásbázisok létrehozásánál kifejezetten hasznos lehet.

- A priori eloszlás alkalmas a meglévő információk vagy annak hiányának jellemzésére.

- A bayes-i szemléletmód egyformán kombinálja az előismereteinkben és a meglévő adatokban lévő információt, megfelelő súlyozást hozva ezzel létre az információk között. Az adatok mennyiségi növekedése befolyásuk növekedését eredményezi.

- Alacsony megfigyelésszám esetén lényeges lehet a posteriori eloszlás ismerete, mivel nem csak a legvalószínűbb eseteket veszi figyelembe, hanem a kevésbé valószínűeket is.

4. fejezet

Klasszikus becslés

4.1. Elméleti hátttér

Fontosnak érzem néhány fogalom[1] ismertetését, hogy a gyakorlati példa folyamatának megértése könnyebb legyen.

4.1.1. Klasszikus pontbecslés

Általánosságban, ha egy X valószínűségi változó eloszlásfüggvénye függ m számú ismeretlen paramétertől:

$$F(x, a_1, a_2, \ldots, a_n)$$

illetve az X-re tartozó n számú mérés kimenetele $X_1, X_2, \ldots, X_m$, akkor az ismeretlen $a_i$állandók becslését a mintaelemek

$$b_i = b_i(X_1, X_2, \ldots, X_n)$$

függvényei, statisztikái segítségével végezzük. A b_i függvényeket az a_i paraméterek statisztikai becslésének nevezzük. Legtöbbször az X valószínűségi változó ismeretlen várható értékét, vagy szórásnégyzetét szoktuk becsülni. A várható értéket :

$$E(X) = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

[1] Lukács Ottó: Matematikai statisztika példatára szolgált a fejezet alapjául.

mintaközéppel becsülhetjük a szórásnégyzetet:

$$D^2(X) = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}$$

korrigált tapasztalati szórásnégyzettel becsülhetjük. Mindkét becslés a nagy számok törvénye miatt megfelelő.

A gyakorlatban sokszor találkozhatunk azzal, hogy a vizsgált X véletlen változó eloszlása ismert (ekkor tudjuk az $F(x)$ eloszlásfüggvényt, vagy $f(x)$ sűrűségfüggvényt), de a sűrűségfüggvény vagy eloszlásfüggvény ismeretlen $T_1, T_2 \dots T_n$ paramétertől függ, amit a statisztikai mintából szükséges becsülnünk.
Ekkor:

$$f(x) = f(x, T_1, T_2 \dots T_n),$$

$$F(x) = F(x, T_1, T_2 \dots T_n)$$

A T_i paraméterek becslését jelöljük az alábbi módon:
$\hat{T}_i (i = 1, 2, \dots, n)$, ahol $\hat{T}_i$ a mintaelemek függvénye (véletlen változó). Az eloszlás egy általunk nem ismert paraméterét közelíthetjük:

- egyetlen számértékkel, mint azt az előbb jelöltük. Az a pontbecslése a mintaelemek valamely $\alpha(X_1, X_2, \dots, X_n)$ függvénye. Maga is valószínűségi változó.

- egy (α_1, α_2) intervallummal, amely nagy valószínűséggel tartalmazza α-t. Ez az intervallumbecslés.
  Ekkor: $P(\alpha_1 \leq a < \alpha_2) = 1 - \varepsilon$
  ahol ε egy nullához közeli alacsony valószínűség. Az $1-\varepsilon$ a fenti (α_1, α_2) megbízhatósági (konfidencia-)intervallumhoz tartozó valószínűség százalékos értéke: $100(1 - \varepsilon)$ százaléka a biztonsági szint. Az α_1, α_2 is mintaelemek függvénye.

A pontbecslés módszerei: Több olyan módszer is ismert, amivel az általunk nem ismert $T_i = T_i(X_1, X_2, \ldots, X_n)$paraméter "jól" konstruálható.

- **Momentumok módszere:**
  Egy olyan eljárás a paraméterek becsléseinek megkonstruálására, ami a minta momentumainak és az eloszlás momentumainak megfeleltetésén alapul. Az esetek nagy részében egy könnyen megoldható egyenletrendszert kapunk az ismeretlen paraméterekre. R. A. Fisher kimutatta, hogy asszimetrikus eloszlások esetén ez a módszer nem annyira hatékony.

- **Maximum - likehood módszer:**
  Ennél a módszernél tudjuk a sokaság eloszlását, de nem tudjuk az eloszlást jellemző paramétert (paramétereket). A paraméter értékét olyan értékkel becsüljük, amiknél a minta bekövetkezése lenne a legnagyobb valószínűségű. A maximális valószínűséget a minta valószínűségét megadó likelihood-függvény maximumával vagy logaritmusának maximumával keressük meg. A likelihood függvény egy T paraméter esetén:

  1. **Diszkrét esetben:**

  $$L(X_1, X_2, \ldots, X_n, T) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) =$$

  $$= P(X_1 = x_1)P(X_2 = x_2) \ldots P(X_n = x_n), \text{ha függetlenek}$$

  Szorzat helyett sokkal könnyebb lehet összeget kezelni, így gyakran a következő függvényt tekintik a likelihood függvénynek:

  $$lnL = \sum_{i=1}^{n} lnP(X_i = x_i)$$

  2. **Folytonos esetben:** Mivel minden pont felvétele az n dimenziós térben 0-a, így azt a valószínűséget kell maximalizálni, hogy a pont az $x_1, x_2, \ldots, x_n)$ pont környezetébe, azaz

  $$x_1 \leq X_1 \leq x_1 + \triangle x_1, \ldots, x_n \leq X_n \leq x_n + \triangle x_n$$

az n dimenziós térbe essen. Ennek a valószínűsége:

$$f(x_1,T)f(x_2,T)\dots f(x_n,T)\triangle x_1,\triangle x_2\dots\triangle x_n.$$

és ez ott maximális, ahol:

$$L(X_1,X_2,\dots,X_n,T)=f(x_1,T)f(x_2,T)\dots f(x_n,T)$$

függvény maximális. Ezt vagy ennek logaritmusát :

$$lnL=\sum_{i=1}^{n}lnf(x_i,T)$$

függvényt tekintik a likelihood-függvénynek. Így ennek megfelelően a

$$\frac{dL}{dT}=0 \text{ vagy a } \frac{dlnL}{dT}=0$$

egyenlet a likelihood egyelet.

- **Legkisebb négyzetek elve:**
  A módszer lényege, hogy az eltérések négyzetösszegét igyekszik minimalizálni. Ezt a módszer főleg a korreláció és regresszió elemzésnél használandó. Mivel a regresszió elemzés nem tartozik a szakdolgozatom témájához, így ezt a módszert nem fogjuk részletesebben bemutatni.

4.1.2. Intervallumbecslés

Eddig egy eloszlás egy ismeretlen T paraméterét egyetlen $\hat{T}$ számadattal becsültük. Az így kapott pontbecslés helyett a most bemutatásra kerülő intervallumbecslést több információt tartalmazandónak érezhetjük. A pontbecslés egyetlen értéket ad, amelynél nem tudható, hogy hány százalék valószínűséggel fog bekövetkezni.
Az intervallumbecslésnél viszont:

- szem előtt tartjuk, hogy $\hat{T} = \hat{T}(X_1, X_2, \ldots, X_n)$ maga is véletlen változó, amely értékei T paraméter valódi értéke körül szóródnak.
- ha ismerjük $\hat{T}$ eloszlásfüggvényét, meg tudjuk mondani, hogy mekkora valószínűséggel tartalmazza (α_1, α_2) intervallum T valódi értékét.

Gyakran nincs szükségünk az eloszlásfüggvény valamely paraméterének pontos becslésére. Elegendő számunkra, ha a mintaelemek két statisztikai függvényével megadunk egy olyan intervallumot, mely előre meghatározott valószínűséggel tartalmazza az ismeretlen paramétert.
Legyen az X valószínűségi változó eloszlásfüggvénye $F(x, a)$, és az a ismeretlen paraméter értéke valamely (a_1, a_2) intervallumba essen. Nézzük az X változó n elemű $X_1, X_2, \ldots, X_n$ mintáját. Ha egy adott $1-p$ valószínűséghez létezik $\alpha_1 = \alpha_1(X_1, X_2, \ldots, X_n; p)$ és $\alpha_2 = \alpha_2(X_1, X_2, \ldots, X_n; p)$ statisztikai függvény, melyre a $(\alpha_1; \alpha_2)$ intervallum $1-p$ valószínűséggel tartalmazza az a nem ismert állandót, akkor (α_1, α_2) az a paraméter konfidencia-intervalluma $1-p$ megbízhatósággal, ahol p egy 0-hoz közel eső alacsony valószínűség.
Nézzünk példaként egy konfidencia-intervallum becslést a normál eloszlás m várható értékére és σ szórására:

Az m várható értéket $\hat{X}$ mintaközéppel becsülve, $\hat{X}$ hasonlóképp normál eloszlású, m várható értékű, és $\frac{\sigma}{\sqrt{n}}$ szórással. ekkor az $\frac{\hat{X}-m}{\sigma/\sqrt{n}}$ változó már szimmetrikus a tengelyre, és standard normális eloszlású.

Mivel szimmetrikus és monoton csökkenő, ezért érdemes a megbízhatósági intervallumot a 0-ra szimmetrikus $-u_\varepsilon$ és $+u_\varepsilon$ határokkal keresni, úgy:

$$P\left(-u_\varepsilon \leq \frac{\hat{x} - m}{\sigma/\sqrt{n}} \leq u_\varepsilon \right) = 2 * (1 - \Phi(u_\varepsilon)) = 1 - \varepsilon$$

legyen, ahol ε egy előre megadott kis valószínűség. Az egyenlőtlenség rendezésével :

$$P\left(\hat{x} - u_\varepsilon \frac{\sigma}{\sqrt{n}} \leq m \leq \hat{x} + u_\varepsilon \frac{\sigma}{\sqrt{n}} \right) = 1 - \varepsilon$$

Ekkor az $\hat{x}$-ra szimmetrikus

$$\left(\hat{x} - u_\varepsilon \frac{\sigma}{\sqrt{n}}; \hat{x} + u_\varepsilon \frac{\sigma}{\sqrt{n}} \right)$$

konfidencia-intervallum az eseteknek átlagosan $1-\varepsilon$*100 százalékában tartalmazza az ismeretlen m várható értéket, ha az intervallumbecslést elég sokszor elvégezzük. Ekkor nem egy becslés esik bele ekkora eséllyel, hanem sok becslésből ennyi százaléka belekerül az intervallumunkba.

Megjegyzések:

- A biztonsági szint alacsonyabbra vételével konfidencia intervallum keskenyedik, ha magasabbra vesszük szélesedik. Ez egészen nyilvánvaló, mert ha az esetek nagyobb százalékában tartalmazó intervallumot szeretnénk megkapni, az egyértelműen szélesebb intervallum is lesz(a változó nagyobb intervallumba nagyobb eséllyel esik).

- Ha a várható értékről semmit nem tudunk célszerű az átlagra szimmetrikus intervallumot keresni. Azonban a gyakorlatban nagyobb valószínűségű, hogy a várható érték nagyobb mint az átlagérték. Ilyenkor célszerűbb egy az értékek átlagára nem szimmetrikus konfidencia-intervallumot keresni.

Nyilvánvalóan akár csak a Bayes módszernél, úgy a klasszikus megközelítésnél is a különböző szituációk határozzák meg a modellt. Más konfidencia-intervallum becslés adódik exponenciális eloszlást feltételezve, mint amit mondjuk a binomiális eloszlással kapnánk. Itt is nagy szerepe van annak, hogy mit is kívánunk becsülni.

4.2. Gyakorlati példa

Itt is ugyanazt a példát fogjuk fölhasználni, mint a Bayes módszer bemutatásánál (3.3. Gyakorlati példa). Kaptunk egy lottósorsolási eredményt és kíváncsiak vagyunk, hogy hány számból húzzák. Ne feledjük a matematikai modell itt is ugyanaz.
Azaz a keresett eseményvalószínűség az alábbi lesz: $\frac{\binom{x_7-1}{6}}{\binom{N}{7}}$.

A klasszikus becsléselmélet megpróbálja elkerülni a szubjektívnek tartott állapot-valószínűségeket. A becslés folyamata így az alábbi:

4.2.1. Pontbecslés

Pontbecslésnek azt az N értéket fogjuk választani, amelyre a keresett eseményvalószínűség a legvalószínűbb. Ekkor az $x_7 = N$ fogjuk kapni, mivel a már meghatározott direkt $\frac{\binom{x_7-1}{6}}{\binom{N}{7}}$ valószínűségek között ez a maximális. Ez a becslés nem torzítatlan, mert várható értékben nem adja meg a valódi értéket, hasonlóan a Bayes-i pontbecslésnél. Így tehát a klasszikus pontbecslésünk értéke is $x_7 = 49$ lesz.

4.2.2. Intervallumbecslés

Most egy becsült intervallumot szeretnénk meghatározni. Keressünk rögzített N mellett egy olyan tartományt, ahová $(1 - \alpha)$ valószínűséggel x_7 esni fog. Mivel a legvalószínűbb rögzített N mellett az $x_7 = N$, és ettől fogva monoton csökken, ezért a becsült intervallum egy $[N(\alpha), N]$ típusú intervallum lesz, ahol a bal végpontot úgy kell megadni, hogy a lehető legnagyobb érték legyen (ekkor teljesül a becsült intervallumra vonatkozó minimalitási tulajdonság), amire még teljesül:

$$\sum_{Z=N(\alpha)}^{N} \frac{\binom{Z-1}{6}}{\binom{N}{7}} \geq 0,95$$

A törtek nevezője állandó, így elegendő csak a számláló értékeit összegezni. Tudjuk, hogy

$$\sum_{Z=5}^{N} \binom{Z-1}{6} = \binom{N}{7}.$$

Az összeg így egyenlő lesz:

$$\frac{\binom{N}{7} - \binom{N(\alpha)}{7}}{\binom{N}{7}} = 1 - \frac{\binom{N(\alpha)}{7}}{\binom{N}{7}}$$

Ekkor a keresett $N(\alpha)$ az alábbi egyenlőtlenséget kell, hogy kielégítse:

$$\frac{\binom{N(\alpha)}{7}}{\binom{N}{7}} \leq 0,05.$$

A binomiális együtthatókat kifejtve a következő egyenlőtlenség adódik:

$$N(\alpha)(N(\alpha)-1)(N(\alpha)-2)(N(\alpha)-3)(N(\alpha)-4)(N(\alpha)-5)(N(\alpha)-6) \leq$$

$$\leq 0,05 * N(N-1)(N-2)(N-3)(N-4)(N-5)(N-6)$$

Ha fordítva nézzük és $x_7 = N(\alpha)$ ismert a konfidenciaintervallum már könnyedén megkapható, csak most az egyenlőtlenségből N keresett. Ekkor a fenti egyenlőtlenség alábbi formáját használjuk:

$$20 * x_7(x_7-1)(x_7-2)\dots(x_7-6) \leq N(N-1)(N-2)\dots(N-6)$$

N értékének megadásához egy 7-ed fokú egyenlet gyökét kell megadnunk. Használjuk föl a 2012.05.19-ei lottóeredményt.

$$2, 17, 22, 30, 35, 49|45$$

Ekkor:

$$20*49*48*47*46*45*44*43 = 8658778867200 \leq N(N-1)(N-2)\dots(N-6)$$

Amelynek 7-ik gyökét vonva a $70,5$ kapunk. Itt is fölfelé kell kerekítenünk, mert $70,5$-t 7-edik hatványra emelve nem adja vissza az értéket. Emiatt az intervallum az alábbi lesz: $[49, 71]$.

4.3. Klasszikus megközelítés szignifikancia tesztjének kritikája

A szignifikancia teszt[2] lényege a következő: adott egy hipotézis valamely ismeretlen paraméterről. Szeretnénk a kisérletezéssel szerzett adatokból eldönteni, hogy hipotézisünk megfelelő volt-e.

Fisher egyik legfontosabb ötlete volt a véletlenszerű mintavétel fogalmának kialakítása. A valószínűséget, mint véletlen jelenséget kezelte. A fő irányvonala az volt, hogy a hipotézis helyességének feltételezése mellett meghatározható egy olyan tartomány, ahová a kisérlet eredményének nagy valószínűséggel esnie kell. Ha nem esik bele a tartományba akkor azt mondhatjuk, hogy szignifikáns az eredmény és a hipotézisünk elvethető (asszimetrikusan).

Itt az elvethetőség alatt nem azt értjük, hogy a hipotézisünk hamis lenne, hisz semmi sem kényszeríthet minket arra, hogy egy hipotézist hamisnak tartsunk. Valami mellett az értékéért döntünk, azaz olyan cselekvést választhatunk, amely magával vonza az érték (anyagi, vagy ideális) megszerzését. Ezzel ellentétben egy kijelentéshez rendelt valóságérték - igaz vagy hamis volta -nem érték, hanem egy ítélet, amely egy másik fogalmi kategória. Tehát a példák egy osztályára az elvetni és elfogadni kifejezés hibás és nincs mit eldönteni. Ilyen például az, hogy a születendő gyermekek fiú-lány aránya 1:1 $= H_0$ hipotézis (ugyanakkora valószínűséggel születik lány, mint fiúgyermek). Itt nem értelmezhető az "elvetni" kifejezés, ellentétben a termelés minőségi ellenőrzésénél.

Ha az eredmény nem szignifikáns , akkor "értéktelen", és nem tudunk semmit sem mondani. A hipotézis helyességére semmilyen információnk nincs. A másik probléma az, hogy a hipotézis fennállásának valószínűségét szeretnénk megtudni, de ez a gyakoriság keretében értelmezhetetlen (emiatt nőtt a Bayes-statisztika utáni érdeklődés).

[2]Ennek a fejezetnek Dr. Vancsó Ödön doktori disszertációjának 3. fejezete és Wickmann: Bayes statisztika című könyve szolgált forrásként

A szignifikancia-tesztnél a fordított kérdésre kapunk választ, nem arra amire tulajdonképp kíváncsiak vagyunk. Azaz, ha a hipotézis igaz, akkor az eredmény mennyire valószínű vagy valószínűtlen. Bayes-statisztikával pontosan a kérdésünkre kapunk választ.

Emiatt nagyon lényeges elem a hipotézis megfogalmazása és a szignifikancia-szint fogalma. Fisher a hagyományos 5 százalékról beszél, ami időnként módosul 1 százalékra. Nyilván gyógyszerek esetében az 5 százalékos hibahatár elfogadhatatlan lenne. A valóságban semmilyen logikai oka sincs annak, hogy az 5 százalékos értéket használjuk. Bármilyen értéket választhatnánk tetszés szerint.

Mint a történeti háttérben is megemlítettük Neyman és Pearson átdolgozta Fisher elméletét és szignifikancia-szint helyett bevezették az első és másodfajú hibát, amit a teszt erejének megadására használtak.

Fisher a szignifikancia szintet a hipotézis "valószínűségeként" szerette volna értelmezni, ami azonban csak a Bayes-i keretek között volna lehetséges. Élete során módosított álláspontján. Míg 1935-ben[3] azt tartotta, hogy a szignifikancia- szintet előre kell rögzíteni, a teszt lefolyása előtt. Később már, 1956-ban[4] az volt a véleménye, hogy a teszt elvégzése után számolandó a pontos szignifikancia-szint. Ekkor már a szignifikancia-szint nem a teszt sajátosságaként jelenik meg, hanem az adatok tulajdonságaként értelmezendő.

Neyman és Pearson két fontos fogalommal bővítette ki a Fisher-féle elméletet és ezzel más alapokra is került a konstrukció. Az asszimetria, mint probléma megoldásaként létrehozták az alternatív hipotézis fogalmát. Ekkor már nem csak nullhipotézisünk van, hanem ezzel ellentétes egy vagy több alternatív hipotézisünk is. Majd definiálták a hibafogalmakat. **Elsőfajú hiba** az, mikor a nullhipotézist elvetjük pedig igaz lett volna. Ez a Fisher-féle szignifikancia-szint másfajta megfogalmazása. **Másodfajú hiba** az, mikor megtartjuk a nullhipotézist pedig az alternatív hipotézist kellett volna elfogadnunk.

[3] R.A.Fisher(1935) The Design of Experiments

[4] R.A. Fisher(1956) Statistical methods and scientific interence

Ez akkor válik számolhatóvá, ha konkrét alternatív hipotézisünk van.

Fontos megjegyeznünk, hogy az első fajú hiba valószínűségének csökkentésével-növelésével a másodfajú hiba valószínűsége is változik. Ha csökkentjük, akkor a másodfajú hiba sokkal nagyobb mértékben fog megnőni, mint amennyivel az eslőfajú hiba valószínűségét megváltoztattuk. Ez csak akkor nem történik így, ha az első fajú hiba csökkentésével a kísérletszámot megnöveljük.

A Neyman, Pearson konstrukciónál van egy hipotézisünk, és egy ezzel szemben álló alternatív hipotézisünk. Ezt a kísérlet alapján kell eldöntenünk. A két hiba ekkor a téves döntések valószínűsége. Azaz, mikor helytelenül elfogadjuk, vagy elvetjük az egyik, vagy másik hipotézist. A hipotézisvizsgálathoz hozzárendelt döntési szabály egy tévedési költség, amely megmutatja, hogy melyikből származik nagyobb kár a hipotézis téves elvetéséből , vagy a hipotézis téves megtartásából.

A Fisher - Neyman, Pearson szemléletmód különbözőségének fő oka az, hogy míg Fisher a filozófiai hátteret, addig Neyman, Pearson páros az alkalmazhatóságot vizsgálta. A Neyman, Pearson féle fogalmi kör másik lényeges fogalma a konfidencia-intervallum.

A Fisher-féle megbízhatósági-intervallumnál keressük az ismeretlen paraméter értékeire egy megbízhatónak tekinthető tartományt. Azt viszont nem tudjuk megítélni, hogy tényleg ebben az intervallumban található-e a paraméter. Erről nem is beszélhetünk, mert a hipotézis igaz, vagy hamis volta nem a véletlentől függ. Így nem lehet hozzá gyakoriségi szempontból valószínűséget párosítani. Valójában információhiány áll fenn, és nem véletlen folyamat. Csak azt mondhatjuk, ha ebből a tartományból származik az adott érték, akkor nagy valószínűséggel hihető a kisérlet eredménye.

A Neyman-Pearson féle konfidencia-intervallum azon paraméterek összesége, amelyre nem szignifikáns az eredmény. Tehát a kapott érték nem vethető el, ha ekkora lenne a paraméter valódi értéke. Azaz itt a paraméter esik bele egy ismeretlen tarományba. Ez is a Fisher-féle szignifikancia teszt egy másfajta megfogalmazása.

A Fisher féle elméletben a nem szignifikáns eredményből nem lehet következtetést levonni, mert ilyenkor csak az elvethetőséget nem tudjuk bizonyítani, nem pedig a hipotézis igaz voltát. Ekkor közömbös eredményekhez jutottunk, amelyből semmiféle következtetést nem vonhatunk le.

Neyman és Pearson mégis kitüntetett tartománynak nevezik azon értékek halmazát, amelyet nem lehet elvetni. Itt tehát nem azt vizsgáljuk, hogy egy rögzített intervallumba mekkora eséllyel esik a becsülni kívánt paraméter, hanem azt, hogy a véletlen intervallumok mekkora eséllyel fednek le egy bizonyos értéket. A bayesianus közelítésben a szignifikancia-szint a hipotézis igaz voltának valószínűsége.

A Neyman, Pearson elméletei az 50-es, 60-as években váltak ismertté. Ekkor alakult ki a mai klasszikus statisztika, ami a Neyman-Pearson-Fisher elméletek összefonódásával jött létre. A tankönyvek többségében nem tesznek említést az egymással szembenállásukról. Sőt összemosva kezelik a kétféle közelítési módot.

5. fejezet

Összehasonlítás

5.1. Gyakorlati példa eredményeinek összevetése

Felmerülhet a kérdés, mikor is esik egybe a Bayes-elemzés eredménye a klasszikuséval, és mikor kapunk attól eltérőt. Elméletben a válasz az, hogy az eredmények nem összehasonlíthatóak, így nem is juthatunk azonos eredményre a kétféle módszer alkalmazásával. Ez még akkor is igaz, ha a számszerű eredmények megegyeznek. Fontos megjegyeznünk, hogy az eredmények numerikus egyenlősége, csak speciális esetben áll fönt. Ilyen eset, ha a mintaeloszlás az egyenletes eloszláscsaládhoz tartozik. Ez egy szükséges feltétele a numerikus egyenlőségnek, de nem elégséges. Ahhoz még föl kell tennünk, hogy a priori eloszlás $f(p) = konstans$. Az ettől másabb priori eloszlás módosítaná a posteriori eloszlásunkat. Nézzük meg az alábbi hipotézis párokat: Klasszikusan: $H_0 : p = p'$ pont hipotézis áll szemben az összevont $H_A : p > p'$ hipotézissel szemben.
Bayes-módszernél: a $H_1 : p \leq p'$ és $H_2 : p > p'$ összevont hipotézispárok állnak szemben egymással. Ekkor már látható, hogy a két módszer hipotézisei is eleve különbözőek. Fontos felsimernünk, hogy még a numerikusan egyenlő eredmények is másként értelmezendőek klasszikus és Bayes-féle modellben.

A fenti lottós példában láthattuk, hogy a Bayes-i és a klasszikus pontbecslésre ugyanazt az eredményt kaptuk. Immár tudjuk, hogy ebben a példában teljesültek a feltételek, amik az egyenlőséghez vezetnek.

Az intervallum becslésnél a Bayes féle megoldás egyenlete az alábbi volt:

$$T(T-1)(T-2)\dots(T-5) \geq 20(x_7-1)(x_7-2)\dots(x_7-6) \qquad (5.1)$$

ahol a legkisebb T értékét kerestük, amire a fenti egyenlőtlenség teljesül. A keresett 0,95-intervallum becslése a $[x_7, T]$ tartomány lett. Míg a klasszikusnál:

$$N(N-1)(N-2)\dots(N-6) \geq 20 * x_7(x_7-1)(x_7-2)\dots(x_7-6) \quad (5.2)$$

A becslés az $[x_7, N]$ intervallum lett. Itt már egyértelűen észrevehető mért nem egyezhetnek numerikusan az egyenletek. Annak ellenére sem, hogy a speciális feltételek fönnállnak. Azaz egyenletes eloszlást feltételezünk és K tart a végtelenbe.

Az (5.1)-es egyenlőtlenség megoldása mindig nagyobb lesz, mint az (5.2)-esé. Emiatt a klasszikus közelítésnél a $0,95$-ös konfidencia-intervallum mindig szűkebb lesz. Ha feltesszük $T = N$ igaz voltát, akkor láthatóvá válik, hogy az (5.1)-es egyenlőtlenség biztosan teljesül, ha az (5.2)-es is. A Bayes tartomány mindenképp bővebb lesz. Ezt láttuk a gyakorlati példánkban is, hisz a Bayes-féle becslésünkre a $[49; 75]$-ot , míg a klasszikusra a $[49, 71]$ konfidencia-intervallumot kaptuk.

5.2. Bayes-i és a klasszikus megközelítés összehasonlítása

A Bayes-i szemléletmódban minden lehetséges információt fölhasználunk, majd ezeket kombináljuk a mintavétel eredményével. A lényeg a priori eloszlás feltételezése, mivel ennek megválasztásától függ az eredményünk. Ha rosszul választjuk meg a priori eloszlásunkat,- például úgy, hogy valamely értékhez 0 valószínűséget párosítottunk holott ettől lényegesen nagyobb a valószínűsége- akkor a valóságostól eltérő, helytelen eredményt is kaphatunk. Emiatt nagyon fontos szerepe van a priori eloszlásnak. Amiből a fenti eljárásokat alkalmazva posteriori eloszláshoz jutunk, mely segít meghatározni az intervallumokat. Ellenben Szergej Natanovics Bernstein és Richard Mises 1920 előtt már bebizonyították, hogy a Bayes tétel ismételt alkalmazásával a posteriori eloszlás sorozata a tényleges megoldáshoz konvergál bizonyos feltételek mellett. Így aszimptotikusan nincs jelentősége a priori eloszlás megválasztásának, persze a feltételek betartásával. A Bayes statisztika lehetővé teszi a szubjektív vélemények kezelését, szubjektív valószínűségek fölhasználásával. A bayes-i felfogásnál a paraméter valószínűségi változó, és a külső információk lényeges szerephez jutnak. Csak egyetlen minta van. A Bayes statisztika tartalmazza speciális esetként a klasszikus statisztikát, azaz ha nincs információnk egyenletes eloszlást alkalmazunk, így ugyanarra az eredményre jutunk mindkét esetben. Ma már tudjuk, hogy a Bayes-tétel átfogalmazása mindenféle statisztikai modellre alkalmazható. Az egyszerűsége és egységessége miatt egyre népszerűbbé és szélesebb körben alkalmazhatóvá válik. A klasszikus statisztika nem képes a szubjektív vélemények kezelésére. A valószínűségeket objektívnek tekinti, a véletlent létrehozó rendszer egyedi tulajdonságának (pl. lottó). Így a mért tapasztalataink nem használhatóak föl, vagy csak nagyon kis mértékben. Azaz nincs nagy jelentőségük a külső információknak. A klasszikusnál is csak egyetlen minta van, de ott feltételezzük az ismételt mintavétel lehetőségét, hiszen az eredmény értelmezése pont több kísérletre vonatkozik.

6. fejezet

Alkalmazási területek

A Bayes-i megközelítési módszer egyre elterjedtebb és rendkívül széles körben alkalmazható. Néhány alkalmazási terület:

- Gazdasági területeken például az idősorok elemzésénél használható föl. Bármilyen legalább kétváltozós modellnél alkalmazható a Bayes-i elemzés. Így például segíthet:
  - Üzleti ciklusok vizsgálatánál. Azaz a gazdasági teljesítmény hullámzásai mennyire hasonlítanak egymásra egyes országok esetében. Észre vehető-e valamilyen kapcsolat közöttük.
  - Előrejelzési idősorok vizsgálatánál. Megtudhatjuk mennyire jelez időben előre az adott rendszer vagy esetleg megkésve mennyire követi az adott gazdasági változó az előrejelzéseket.
  - Fiskális és monetáris politika csatornáinak vizsgálatára is alkalmas. Például a fiskális intézkedés jövedelem megszorításaira hogyan reagál a lakossági fogyasztás. Vagy egy esetleges monetáris megszorítás mikor és milyen hatással lesz az inflációra.
- Orvoslás területén is nagyon hasznos lehet. Szintén idősorelemzés keretén belül lehetőséget nyújt jobb orvosdiagnosztikai eszköz létrehozásában. A már meglévő, korábbi mintákat újraelemezve olyan részletekhez juthatunk, melyek segíthetnek a biológiai folyamatok jobb meg-

értésében. Korábban felismerhetővé válnak az elváltozások, így hamarabb megkezdhető lesz a kezelés. 1998 óta a Leuveni egyetemen folyó petefészekrák- diagnosztikában alkalmazzák. A 2000-20002 között alakult IOTA (Intenational Ovarium Tumor Analysis)nemzetközi konzorcium vitte tovább a kutatásokat, mely a legjobb petefészekrák kutatókat tömöríti magába.

- Élelmiszertechnológiai problémák kezelésére is alkalmas. A fogyasztók számára egyre nagyobb szerepet kap az, hogy megfelelő minőségű terméket vásároljanak. Így fontossá vált, hogy egészségre ártalmas, toxinokkal mérgezett ételek ne kerülhessenek a boltok polcaira. A génmanipulált termékek megjelenésével sok fogyasztóban felmerült az ezek fogyasztásával vállalt kockázat is. A kockázatbecslés segítségével meghatározzák a káros hatást okozó veszélyforrásokat, majd jellemzik a következményeket. A hagyományos toxikológiai vizsgálat során 50-100-szorosnál mérik a lehetséges hatást, de ez génmanipulált termékek esetében lehetetlen. Az élelmiszerek nagyon komplex, inhomogén összetételű anyagok. Ezért vizsgálatuk során sok a bizonytalanság, amit a klasszikus módszerek nem képesek kezelni. Viszont a Bayes statisztika képes a fenti probléma megoldására (Bayes hálókat alkalmaznak, ami a többdimenziós paraméterek közötti összefüggés könnyebb kezelhetőségére szolgál). Fontos megjegyeznünk, hogy a Bayes módszert nem az általános esetek vizsgálatánál alkalmazzuk, hanem ha valami "extrém" esetre vagyunk kiváncsiak.

- Közvéleménykutatásnál is fontos szerepe van. A közvéleménykutatók szeretnék megtudni a populáció megoszlását egy adott kérdéssel kapcsolatban egy meglévő minta segítségével. Ez viszonylag egyszerűen kiszámolható pont és intervallumbecslés fölhasználásával.

- Statisztikai elemzéseknél is alkalmazható. Tekintsük az alábbi gyakorlati feladatot:

6.1. Gyakorlati példa

Dieter Wickmann Bayes statisztika című könyvében a 144. oldalán található 6.11.2 feladat 2-es részfeladatának módosítását fogom bemutatni. A feladat számolását én végeztem, valósághű adatokat tartalmaz.

A Központi Statisztikai Hivatal honlapjáról lekérdezett adatokból tudjuk, hogy 2011 első félévében a halálos autóbalesetek száma 237, míg 2011 július-augusztusában 114 volt Magyarországon. Tegyük fel, hogy a rendőrség újfajta felvilágosító programot vezetett be az autósok számára 2011 júniusában, hogyan tudnák a baleseteket megelőzni. A program hatékonyságát kívánják felmérni. Azaz tudni szeretnék növekedett-e a baleseti ráta.

Ahhoz, hogy megtudjuk növekedett-e a ráta, két hipotézist kell összehasonlítanunk.
Legyen:
H_1="baleseti ráta változatlan"
H_2="baleseti ráta növekedett"
Fogalmazzuk át a feladatot egy olyan modellre, amelynek a megoldási folyamata ismert. A D.Wickmann: Bayes statisztika 140 oldalán ismertetett urnamodellre és megoldási folyamatára visszavezethetjük.

Mivel a baleseti ráta változására vagyunk kiváncsiak, vegyük úgy, hogy a 2011 július-augusztusi időszakában a balesetek száma változatlan maradt. Ez azt jelenti, hogy a halálos balesetek $p' = \frac{1}{4}$ része az utolsó félévre esik.(2 hónapot fixáltunk a 8 hónapból, így $\frac{2}{8}$ része változatlan a hónapoknak)
Így a két szemben álló hipotézisünk:
$H_1 : p = p'$
$H_2 : p > p'$

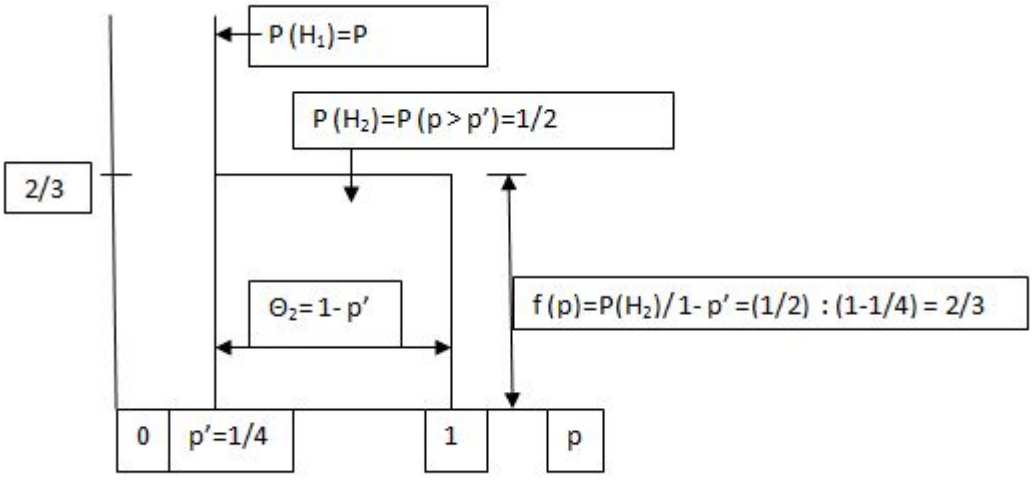
Ekkor legyen a "fehér golyó" a "baleset 2011 július-augusztusában", és a "fekete golyó" a "baleset 2011 első félévében". Tehát az alapmodellünk: Van két urnánk U_1 és U_2, amik fekete és fehér golyókat teszünk. Az U_1 urnában a fehér golyók aránya p megegyezik p'-vel. Az U_2 urnáról csak annyit tudunk, hogy $p > p'$.Tehát U_1-ben $p = p' = \frac{1}{4}$, U_2-ben $p > p'$ fehér golyó van. Szükségünk lesz $P(U_j)$ valószínűségre, ha n=237+114=351 golyó közül 114

fehér. Ahhoz, hogy kideríthessük, melyik urnáról van szó, Bernoulli féle mintát kell vennünk n-szeri, itt most 351 húzással.

Így eljutunk egy θ diszkrét állapothoz, ami $\theta_1 = p'$ pontból és $\theta_2 =]p', 1]$ intervallumból áll. H_1 esetén feltesszük, hogy az első urnából húzztuk a golyót, H_2 esetén pedig a másodikból. Ekkor szeretnénk megtudni $P(H_1|k)$ és $P(H_2|k) = 1 - P(H_1|k)$ valószínűségeket úgy, hogy épp 114 fehér golyó volt a mintában. Így k=114.

Mivel nem ismerünk semmilyen háttérinformációt, így feltehetjük, hogy a valószínűsége annak, hogy az első urnából húzunk megegyezik a második urnából való húzás valószínűségével. Ekkor $P(H_1) = P(H_2) = \frac{1}{2}$. Mivel minden $p \in \theta$ egyformán valószínű, így $\mathbf{f(p) = c}$ konstans lesz.

A fenti priori feltételezések az alábbi sűrűségre vezetnek:



6.1.1

Itt felhasználtuk azt, hogy az összvalószínűség két lépésre lett bontva. Kezdetben arra keressük a választ, hogy az urnák milyen priori valószínűséggel rendelkeznek.

A válasz: $P(H_1)$, és $P(H_2) = 1 - P(H_1)$. Míg $P(H_1)$-et egyetlen p' pontra összpontosítjuk, és $P(H_1) = P(p = p')$, addig a $P(H_2)$ az $f(p)$ sűrűségintegrálja a θ_2 intervallumon.

A következő lépésben $f(p)$-t keressük. Alkalmazzuk azt, amit már a második urnáról tudunk a kísérlet megkezdése előtt. Már kiszámoltuk, hogy $f(p) = c$ konstans$= \frac{2}{3}$. A feladat megoldása ekkor már következik a Bayes-tételből.

θ_2 folytonos esetére az alábbi képlettel számolunk:

$$P(H_i|k) = \begin{cases} \frac{Z_1}{P(k)}, Z_1 = P(p')P_B(k|n,p') = \frac{1}{2}\binom{n}{k}p'^k(1-p')^{n-k} & \text{ha } i = 1 \\ \frac{Z_2}{P(k)}, Z_2 = \int_{p'}^1 f(p)P_B(k|n,p)dp = c\binom{n}{k}\int_{p'}^1 p^k(1-p)^{n-k}dp & \text{ha } i = 2 \end{cases}$$

Ahol $P(k)$ nevező a mintavétel előtti valószínűsége annak, hogy k fehér golyót húzunk.
Z_1 mintavétel előtti valószínűsége annak, hogy az U_1 urnát választottuk és éppen k fehér golyót húzunk.
Z_2 mintavétel előtti valószínűsége annak, hogy U_2 urnát választottuk és k fehér golyót húztunk.
A $\frac{Z_i}{P_k}$ hányados poszteriori valószínűsége annak, hogy a minta eredménye U_i urna útján jön létre. Mivel most két urnánk van, így $P_k = Z_1 + Z_2$

Az értékeket behelyettesítve a fenti képletbe:

$$P(H_i|114) = \begin{cases} \frac{Z_1}{P(114)}, Z_1 = \frac{1}{2}\binom{351}{114}(\frac{1}{4})^{114}(\frac{3}{4})^{237} & \text{ha } i = 1 \\ \frac{Z_2}{P(114)}, Z_2 = \frac{2}{3}\binom{351}{114}\int_{\frac{1}{4}}^1 p^{114}(1-p)^{237}dp & \text{ha } i = 2 \end{cases}$$

Z_2 úgy is megkapható, hogy :

$$Z_2 = \frac{c}{n+1}(1 - F_\beta(p'|k,n)) = \frac{\frac{2}{3}}{352}\Big(1 - F_\beta(\frac{1}{4}|114,351)\Big) =$$

$$= \frac{\frac{2}{3}}{352}\left(1 - \sum_{i=0}^{237}\binom{352}{i}\left(\frac{1}{4}\right)^{352-i}\left(\frac{3}{4}\right)^i\right)$$

alakra hozzuk. Ezt a formátumot már BINOBETA program, vagy tudományos számológép segítségével már kiszámolhatjuk.

Ha nem ismernénk a BINOBETA programot, és nem rendelkezünk tudományos számológéppel-ahogy én sem -akkor online ingyenesen elérhető a Wolfram Alpha nevű program, amivel könnyedén kiszámolható.

Ekkor:

$Z_1 = 1,6267 * 10^{-4}$

$Z_2 = 18,95 * 10^{-4}$

$P(H_1|114) = \frac{Z_1}{Z_1+Z_2} = \frac{1,6267*10^{-4}}{1,6267*10^{-4}+18,925*10^{-4}} = 0,07915$

$P(H_2|114) = 1 - P(H_1|114) = 0,9205$

Tehát annak a valószínűsége, hogy a baleseti ráta nem változott 7,9 százalék míg a növekedésé 92,05 százalék. Ez alapján nem volt hatékony a program.

Irodalomjegyzék

[1] http://hu.wikipedia.org/wiki/Statisztika (2012)

[2] http://en.wikipedia.org/Bayes'_theorem (2012)

[3] elib.kkf.hu/okt_publ/h_003.pdf
(Dr. Horváth Jenőné: A Bayes- statisztika és alkalmazása)

[4] http://web.kvif.bgf.hu/upload/menu_content/doc/20100224151924L_3cikk.pdf
(Dr. Horváth Jenőné: Bayes- i megoldások élelmiszerbiztonsági problémákra)

[5] http://www.ksh.hu/statszemle_archive/2006/2006_10-11/2006_0-11_966.pdf
(Várpalotai Viktor: Időben változó valós rendű eltolás és becslése; Statisztikai Szemle 84. évfolyam 10-11. szám)

[6] http://doboandor.freeweb.hu/pdfs/a_kovetkeztetesi_tetel_k̃ovetkezmenyei.pdf
(Dobó Andor: A következtetési tétel következményei)

[7] http://www.hiradastechnika.hu/data/upload/file/2005/2005_10/HT_0510-8.pdf
(Millinghoffer András- Hullám Gábor- Antal Péter: Statisztikai adat- és szövegelemzés Bayes-hálókkal a valószínűségektől a függetlenségi és oksági viszonyokig, 2005)

[8] http://www.inf.unideb.hu/valseg/JEGYZET/valseg/node24.htm
(Pere Zsolt: Feltételes valószínűség)

[9] http://www.ksh.hu/docs/hun/xstadat/xstadat_evkozi/e_ods001.html (Központi Statisztikai Hivatal: Személysérüléses közúti balesetek 2008-2012ig)

[10] http://xenia.sote.hu/hu/biosci/docs/biometr/course/concepts/# prob (Bevezetés a biometriába-fogalmak)

[11] http://www.lotteryusa.com/lottery/NY/NYlotto_fYEAR.html (LotteryUSA: New York Results and winning number, 2012)

[12] Dieter Wickmann(2004) Bayes-statisztika; ELTE Eötvös Kiadó, Budapest.

[13] Székely J. Gábor(2004) Paradoxonok a véletlen matematikájában; Typotex könyvkiadó.

[14] Obadovics J. Gyula(1995) Valószínűségszámítás és Matematikai statisztika; Scolar Könyvkiadó.

[15] Prékopa András(1962) Valószínűségelmélet; Műszaki Könyvkiadó.

[16] Vancsó Ödön(2005) Klasszikus és Bayes-i statisztika a matematika didaktikában; Phd disszertáció, Debrecen.

[17] Rényi Alfréd(1966) Valószínűségszámítás; Tankönykiadó, Budapest.

[18] K.A: Ribnyikov(1974) A matamatika története; Tankönyvkiadó, Budapest.

[19] Vincze István(1968) Matematikai statisztika ipari alkalmazásokkal; Műszaki Könyvkiadó.

[20] Antal Éva(2010) Bayes típusú problémák; Bsc-s szakdolgozat, Budapest.

[21] Lukács Ottó(1987) Matematikai statisztika példatár; Műszaki Könyvkiadó, Budapest.

[22] Vincze István(1966) Matematikai statisztika; Tankönyvkiadó, Budapest.

Az internetes oldalak utolsó elérési dátuma: 2012.06.30