

EÖTVÖS LORÁND TUDOMÁNYEGYETEM
TERMÉSZETTUDOMÁNYI KAR

Bajnok Eszter

Matematika BSc

HIPOTÉZISVIZSGÁLAT
ALKALMAZÁSAI EGÉSZSÉGÜGYBEN
ÉS KRIMINOLOGIÁBAN
Szakdolgozat

Témavezető: Backhausz Ágnes egyetemi adjunktus
Valószínűségelméleti és Statisztika Tanszék



Budapest, 2025

Tartalomjegyzék

Köszönetnyilvánítás	3
Bevezetés	4
1. Hipotézisvizsgálatról általánosságban	6
2. Egyszerű és hasznos négymező-teszt	8
3. Hipotézisvizsgálat túlélési függvényekre	12
3.1. Kétmintás teszt	13
3.2. Három vagy több mintás teszt	18
3.3. Trend tesztelése	21
3.4. Kétlépcsős eljárás	22
3.4.1. Számolás menete	23
3.4.2. Szignifikanciaszint és p-érték meghatározása	28
4. Biztonsági hipotézis	33
4.1. Gépjárműlopás	34
4.2. Lakásbetörés	37
4.2.1. Anglia és Wales	38
4.2.2. Egyesült Amerikai Államok	44

Köszönetnyilvánítás

Szeretnék köszönetet mondani témavezetőmnek, Backhausz Ágnesnek a rengeteg segítségéért, idejéért és türelméért. Nagyon hálás vagyok, hogy mindig gyorsan, készségesen válaszolt a felmerülő kérdéseimre.

Köszönöm a családomnak és a barátaimnak is a biztatást és a biztonságot adó háttértámogatást.

Bevezetés

A világban számtalan állítás van, amiről jó lenne eldönteni, hogy nagy valószínűséggel igaz-e. Például ilyen az egészségügyben egy új kezelési módszer vizsgálata: hatásosabb-e a régebbinél. A további fejezetekben több módszer is található példákkal együtt (négymező-teszt, log-rank teszt, kétlépcsős eljárás, ...), melyek segítségével meg tudjuk állapítani, hogy elfogadható-e a feltevésünk. Ezekhez az eljárásokhoz csupán nagy mintaelemszámra és valós adathalmazra van szükség.

A négymező-tesztnél adott két tulajdonság alapján csoportokba rendezzük a megfigyeléseket, és az eljárás megmutatja, hogy van-e szignifikáns összefüggés a két szempont között. Erre kiszámolt példa az egészségügyből: két kezelés hatékonyságának vizsgálata a gyógyulás függvényében, illetve Nagy-Britanniában és Németországban a fiatalokat megtámadó Creutzfeldt-Jakob-betegség azonos intenzitású-e.

A túlélés eloszlását vizsgáló módszereket is ismertetjük (ilyen a log-rank teszt, Gehan-féle súlyozással vett próbastatisztika és a trend tesztelés is), ahol szintén csoportokat hasonlítunk össze, viszont itt speciálisan az életben maradás a fő szempont, ami alapján vizsgáljuk az osztályokat. A példáinkban egy mellrákkal rendelkező páciensek adathalmazát elemezzük: a hormonkezelés, a kor, illetve a tumor fokozata hogyan befolyásolja a túlélést (valós adatokon saját elemzés). Külön kitérünk arra az esetre, amikor a tapasztalati eloszlásfüggvények keresztezik egymást, ehhez a kétlépcsős eljárást használjuk.

Nemcsak az egészségügyben vannak eldöntendő állítások, hanem például a kriminológiában is. Az utolsó fejezetben az autólopás és a lakásbetörés kerül előtérbe. Számos tényező befolyásolhatja a bűncselekmények elkövetését. Az a kérdés, hogy mik azok az intézkedések, amik hatásos védelmet tudnak nyújtani a fenti jogsértések ellen. Ennek megválaszolásához a lineáris regressziót használjuk. Egy számomra elég meglepő eredmény az, hogy a riasztók nem nagyon segítenek megakadályozni a betöréseket, sőt van, hogy

ezek a biztonsági berendezések teszik még vonzóbbá a helyszínt.

1. fejezet

Hipotézisvizsgálatról általánosságban

A hipotézisvizsgálatnak az a célja, hogy el tudjuk dönteni egy állításról (ez a nullhipotézis), hogy az adatok alapján elfogadható-e vagy sem (ebben az esetben az adatok szignifikánsan eltérnek a nullhipotézistől). Azaz nem egy eloszlás ismeretlen paraméter értékére vagyunk kíváncsiak, hanem arra, hogy egy állítás teljesül-e rá. Ha elutasítjuk a nullhipotézist, akkor az állítás megcáfolható statisztikai értelemben, mert a megfigyelések ellene szólnak. Azonban, ha elfogadjuk, akkor az adatok alapján lehetséges, hogy az állítás igaz. A nullhipotézist H_0 -al szokták jelölni. Ezzel szemben pedig az ellenhipotézis vagy más néven alternatív hipotézis van (jele: H_1 vagy H_A).

Az első mondatban elrejtettem még egy ismeretlen fogalmat, ami a szignifikancia. Egy statisztikai próba szignifikanciaszintjét vagy más néven a terjedelmét (level of significance) α -val szokták jelölni. Általában $\alpha = 0,05$, amit önkényesen állapítottak meg, de széles körben, nemzetközileg is elfogadott. A tudományos irodalomban egy eredmény akkor szignifikáns, ha H_0 esetén legfeljebb 5% annak a valószínűsége, hogy a véletlenről van szó ($p \leq 0,05$ -öt fogjuk vizsgálni a későbbiekben).

1.0.1. Definíció. *Egy hipotézisvizsgálati feladatban a p -érték a legnagyobb olyan szignifikanciaszint, ami mellett H_0 -t elfogadjuk.*

Természetesen a statisztikailag szignifikáns eredmények is lehetnek hibásak. Erre van az 5%-os tűréshatár, ami azt jelenti, hogy a hibás eredmény 5%-os kockázatát még megengedik. Ezeket a lehetséges tévedéseket hívják elsőfajú hibának, amikor elutasítjuk H_0 -t, pedig valójában igaz. Ennek a

hibának a valószínűségéből számoljuk a szignifikanciaszintet. Ilyen például az automata tűzjelző, ami akkor riasztózik, amikor semmi sem ég (itt a nullhipotézis az, hogy nincs tűz). Ezen kívül még létezik a másodfajú hiba, ami akkor van, amikor elfogadjuk H_0 -t, de nem igaz a valóságban. Erre pedig példa az előző sémát használva az, amikor a tűzjelző nem jelez veszélyt, pedig ég az épület.

2. fejezet

Egyszerű és hasznos négymező-teszt

Az alábbi fejezet főleg az [1]-es forrás alapján készült. A négymező-teszt segítségével megbecsülhetjük, hogy valamely megfigyelt összefüggés két esemény között mekkora valószínűséggel mondható véletlennek vagy törvényszerűnek. Matematikában függetlenségvizsgálatként ismert ez a módszer. Ehhez 2 szempont alapján osztályokba soroljuk a megfigyeléseket (például szín és forma alapján). Az első szempont (szín) alapján legyen $A_1, \dots, A_r$ (ez teljes eseményrendszer), a második szempont (alak) alapján pedig legyen $B_1, \dots, B_s$. Négymező-tesztnél $s = r = 2$. H_0 legyen az az állítás, hogy a két szempont egymástól független, tehát a megfigyelt eltérés véletlen. H_1 pedig legyen az az alternatív hipotézis, hogy a két szempont összefüggő.

További jelölések:

- N_{ij} : azon megfigyelések száma, amelyekre teljesül A_i és B_j is
- $N_{i.} = \sum_{j=1}^s N_{ij}$ (táblázatban az i. sor sorösszege)
- $N_{.j} = \sum_{i=1}^r N_{ij}$ (táblázatban a j. oszlop oszlopösszege)
- n : összes megfigyelés száma

- f : szabadsági fok, aminek az értéke megegyezik $(r - 1) \cdot (s - 1)$ -gyel (ez azt jelenti, hogy $N_{i\cdot}$ -t és $N_{\cdot j}$ -t rögzítve N_{ij} -k közül $(r - 1) \cdot (s - 1)$ darab szabadon választható, a többi viszont már megvan határozva)

A próbastatisztika a következő:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{\left(N_{ij} - \frac{N_{i\cdot} \cdot N_{\cdot j}}{n}\right)^2}{\frac{N_{i\cdot} \cdot N_{\cdot j}}{n}} \quad (2.1)$$

$$\left(\text{ha } r = s = 2, \text{ akkor } \chi^2 = \frac{n \cdot (N_{11} \cdot N_{22} - N_{12} \cdot N_{21})^2}{N_{1\cdot} \cdot N_{2\cdot} \cdot N_{\cdot 1} \cdot N_{\cdot 2}}\right).$$

χ^2 számlálója a nullhipotézistől való eltérést vizsgálja, mert függetlenség esetén $\mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i)\mathbb{P}(B_j)$, és így $\frac{N_{ij}}{n} \approx \frac{N_{i\cdot}}{n} \cdot \frac{N_{\cdot j}}{n} \Leftrightarrow N_{ij} \approx \frac{N_{i\cdot} \cdot N_{\cdot j}}{n}$.

Ahhoz, hogy tudjunk dönteni a nullhipotézisről, még meg kell határozni a kritikus értéket, amit c_{krit} -tel jelölünk. Ez az érték nem más, mint az f szabadsági fokú χ^2 -próba kritikus értéke α szignifikanciaszint mellett, azaz a χ^2 -eloszlás $1 - \alpha$ kvantilise (táblázat vagy programozás segítségével könnyen meghatározható). Ha $\chi^2 < c_{krit}$ (másképp $p \geq \alpha$), akkor elfogadjuk a nullhipotézist. Ha pedig $\chi^2 > c_{krit}$ (azaz $p < \alpha$), akkor meg elvetjük a nullhipotézist, azaz törvényszerű eltérés van a kategóriák között.

Az egészségügyben a következő példákra használható a fenti négymező-teszt.

2.0.1. Példa. *Az alábbi példa forrása: [2]. Az orvostudományban szeretnék meghatározni, hogy egy meghatározott betegség kezelésénél az új eljárás hatékonyabb-e, mint a hagyományos eljárás. Mindkét típusú kezelésnél 30 beteg eredményét feljegyzik, tehát összesen 60 beteget vizsgálnak. A segédtáblázat a következőképpen néz ki:*

	Sikeres	Sikertelen	Összesen
Hagyományos eljárás	11	19	30
Új eljárás	19	11	30
Összesen	30	30	60

2.1. táblázat. Hagyományos eljárás vs új eljárás

Ebből $\chi^2 = \frac{n \cdot (N_{11} \cdot N_{22} - N_{12} \cdot N_{21})^2}{N_{1.} \cdot N_{2.} \cdot N_{.1} \cdot N_{.2}} = \frac{60 \cdot (11 \cdot 11 - 19 \cdot 19)^2}{30 \cdot 30 \cdot 30 \cdot 30} = 4,27$ és $f = (r - 1) \cdot (s - 1) = 1$.

1 szabadsági fok és 5%-os szignifikanciaszint mellett a c_{krit} 3,84-dal egyezik meg. Mivel a χ^2 értéke ennél nagyobb, ezért H_0 -t elutasítjuk, azaz a 2 eljárás közötti eltérés nem véletlen. Tehát az új eljárás sikeresebb, mint a hagyományos, az orvosoknak megéri az előbbit alkalmazniuk a betegeknél (ezt az információt a táblázatból tudjuk kiolvasni, mert a χ^2 -eloszlás csak azt árulja el, hogy van-e szignifikáns összefüggés, azt nem, hogy mi az iránya).

2.0.2. Példa. Most egy mellrákos páciensek adatait tartalmazó táblázat alapján készítjük el a statisztikánkat (forrás: [3]). Itt azt szeretnénk megnézni, hogy van-e különbség azok között, akiket hormonálisan kezeltek, és azok között, akiket nem kezeltek hormonálisan (H_0 legyen az az állítás, hogy nincs különbség közöttük). Ehhez újra elkészítjük a segédtáblázatot (ezt már python segítségével csináltam, mert nagyon sok adat van - a programkód itt nézhető meg: <https://github.com/Beszt1/szakdoga/blob/main/mellrak.ipynb>):

	meggyógyult	kiújult	a rák/meghalt összesen
hormon kezeltek	152	94	246
nem hormon kezeltek	235	205	440
összesen	387	299	686

2.2. táblázat. Hormonkezeltek vs nem hormonkezeltek

A következő lépés megint a χ^2 és c_{krit} kiszámolása:

$$\chi^2 = \frac{n \cdot (N_{11} \cdot N_{22} - N_{12} \cdot N_{21})^2}{N_{1.} \cdot N_{2.} \cdot N_{.1} \cdot N_{.2}} = \frac{686 \cdot (152 \cdot 205 - 94 \cdot 235)^2}{246 \cdot 440 \cdot 387 \cdot 299} = 4,51$$

és c_{krit} továbbra is 3,84, mert a szabadsági fok 1 és $\alpha = 0,05$. Tehát megint az jött ki, hogy $\chi^2 > c_{krit}$, ami azt jelenti, hogy újra elutasítjuk H_0 -t. Ez azzal ekvivalens, hogy szignifikánsan különbözik a 2 eljárás.

Ezt a példát a következő fejezetben részletesebben is elemezzük (az időpontokat is figyelembe véve).

2.0.3. Példa. Az alábbi példa forrása: [2]. Creutzfeldt-Jakob-betegség 2001-ben, főként a fiatalokat veszélyeztető változatát vizsgáljuk meg. Két területen - Németországban és Nagy-Britanniában - nézték meg az utolsó 3 év adatait. A nullhipotézisünk legyen az, hogy a két országban azonos arányban betegszenek meg a fiatalok. Az alábbi táblázat szépen mutatja az esetek számát külön-külön kategóriánként:

	40 év alatti	40 év feletti	Összesen
Nagy-Britannia	9	198	207
Németország	1	80	81
Összesen	10	278	288

2.3. táblázat. Creutzfeldt-Jakob-megbetegedések összehasonlítása

Észrevehető, hogy az egyik cellában 1-es szerepel, emiatt nem teljesülnek a próba elvégzésére vonatkozó feltételek, de nem volt több adat.

χ^2 -re 1,68 jön ki a fenti képletbe behelyettesítve. A kritikus érték továbbra is 3,84, hiszen a szabadsági fok (1) és a szignifikanciaszint (5%) sem változott. A két értéket összehasonlítva most az jön ki, hogy $\chi^2 < c_{krit}$, ami azt jelenti, hogy elfogadjuk a nullhipotézist, azaz nincs szignifikáns különbség a két országban a fiatalok megbetegedésének arányai között az adatok alapján.

3. fejezet

Hipotézisvizsgálat túlélési függvényekre

Ebben a fejezetben túlélési adatokat vizsgálunk meg különböző eljárások alapján (főleg [4] forrásra támaszkodva, és a példák kódjai az alábbi linken érhetők el: <https://github.com/Beszt11/szakdoga/blob/main/mellrak.ipynb>). Ez azt jelenti, hogy n embernél egyféle eseményt nézünk, ami mindenkinél független azonos eloszlással egy véletlen időpontban következik be, és az eloszlást vizsgáljuk. Túlélésnek azt nevezzük, ha egy embernél még nem következett be az esemény. Alkalmazás szempontjából pedig nagyon hasznos, mert ennek segítségével például különböző terápiás módszerek közül ki tudjuk választani azt, amelyik esetében a legnagyobb a túlélés valószínűsége. Nemcsak kezeléseknél, hanem akár gyógyszerek hatékonyságánál is alkalmazhatóak a következő statisztikai próbák.

A továbbiakban az időt napban fogjuk mérni, és diszkrét változóként tekintünk rá, azaz az egy napon belül bekövetkezett események időpontja megegyezik.

Ahhoz, hogy megértsük pontosabban a módszereket, új fogalmakra van szükségünk.

3.0.1. Definíció. *A túlélési függvény egy adott t időpontban annak a valószínűségét adja meg, hogy az adott egyed a kezdő időponttól kezdve t -ig életben marad (röviden: $S(t) = \mathbb{P}(T > t)$, ahol $S(t)$ a túlélési függvény, t az időtartam, T pedig egy valószínűségi változó, mely az esemény bekövetkezéséig eltelt időt jelöli).*

3.0.2. Definíció. *Ezzel szemben a hazardráta annak a valószínűségét adja meg, hogy egy adott személynél t időpontban bekövetkezik az esemény, feltéve, hogy addig nem következett be.*

3.0.3. Definíció. *A hazard függvény a hazardráta idő szerinti függvénye.*

Az adatok cenzorálása azt jelenti, hogy bizonyos megfigyelések esetében nem ismerjük az esemény pontos időpontját, mert az esemény még nem következett be, vagy az adatgyűjtés befejezése után történt meg.

A $h(t)$ kumulatív hazard függvényt ($h(t) = -\log(S(t))$) képlettel fejezhető ki, ahol $S(t)$ továbbra is a túlélési függvény) különböző módszerekkel lehet megbecsülni. Így a próba elvégzése előtt erős sejtésünk lehet az eredményre.

Például a Nelson-Aalen becslővel a következőképpen néz ki a $h(t)$ függvény becslése:

$$h(t) = \sum_{t_i \leq t} \frac{d_i}{Y(t_i)}, \quad (3.1)$$

ahol d_i a t_i időpontban bekövetkezett események száma (általában halálozások száma), $Y(t_i)$ pedig a t_i időpontban vizsgált egyedek száma. Ez a kumulatív hazard függvényt adja a cenzorált túlélési adatokból. Az utóbbi hozzárendelés azt a valószínűséget adja meg, hogy egy egyed a t időpontban elszenved egy eseményt, feltéve, hogy addig nem szenvedett el (itt általában a halálozás szokott lenni az esemény). A Nelson-Aalen becslő azt mutatja meg, hogy az idő előrehaladtával hogyan növekszik az események bekövetkezésének összesített kockázata.

Egy másik lehetőség a Kaplan-Meier módszer. A túlélési görbe becslése ebben az esetben így néz ki az idő alakulásával:

$$S(t_j) = S(t_{j-1}) \left(1 - \frac{d_j}{n_j}\right), \quad (3.2)$$

ahol n_j az adott eseményt - legtöbbször halált - nem elszenvedő egyedek száma a t_j időpont előtt, d_j pedig az események száma a t_j időpontban. Kezdeti feltétele is van ennek a görbének, ami nem más, mint $t_0 = 0$ és $S(0) = 1$.

3.1. Kétmintás teszt

A kétmintás teszt igazából az egymintás próba kiterjesztése, ahol azt vizsgáljuk, hogy az adott populáció hazardrátája minden $t \leq \tau$ esetében $h_0(t)$

(ez egy adott függvény), azzal az alternatívával szemben, hogy a hazardráta nem minden $t \leq \tau$ esetén $h_0(t)$. Általában τ a vizsgálati idők közül az a legnagyobb érték, amire teljesül, hogy mindegyik mintából még van megfigyelt beteg. A nullhipotézis teszteléséhez a megfigyelt és a várható hazardráták súlyozott különbségeinek összegét hasonlítjuk össze.

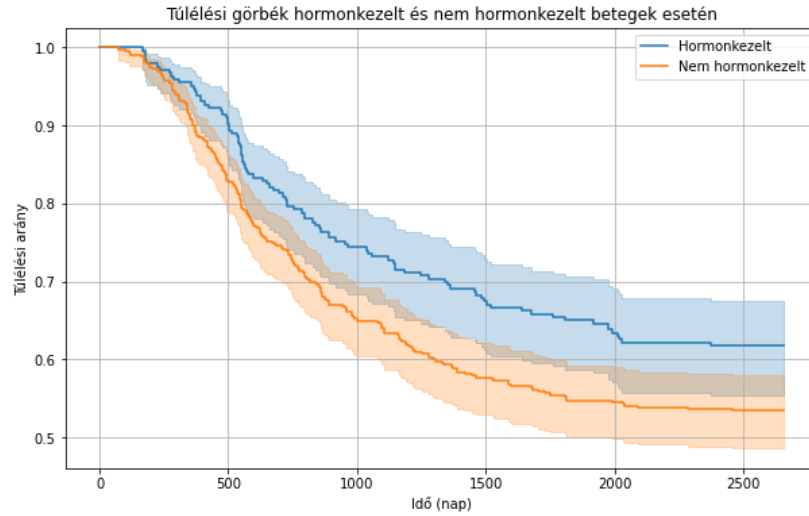
Ha már 2 mintánk van, akkor $H_0 : h_1(t) = h_2(t) \forall t \leq \tau$ esetén és $H_A : h_1(t) \neq h_2(t)$. Ezen felül kell választani a statisztikánkhoz egy pozitív súlyfüggvényt (jele: $W(t)$), amire teljesül, hogy $W_j(t_i) = 0$, ha $Y_{ij} = 0$, ahol Y_{ij} a j . csoportban t_i időpontban vizsgált egyedek száma. A legegyszerűbb eset, ha azonosan 1-nek vesszük a súlyfüggvényt - ez a log-rank tesztet fogja adni. Ezen kívül persze számtalan másik súlyfüggvény is létezik - például Gehan-féle súlyozás, Tarone-Ware-féle súlyozás, Peto-Peto-féle súlyozás A súlyfüggvény meghatározása után még szükségünk van egy segéd táblázatra, ami segítségével z értékét könnyen ki tudjuk számolni (a 3.1.1 példában részletesen be van mutatva). Egyszerűen csak az utolsó kettő oszlop értékeit kell szummázni D -ig (ahol D az utolsó bekövetkezett esemény időpontja), mert ez lesz a z számlálója, illetve nevezője a négyzetgyök alatt

$$\left(z = \frac{\sum_{i=1}^D W(t_i) \left(d_{i1} - Y_{i1} \frac{d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^D W(t_i)^2 \frac{Y_{i1}}{Y_i} \left(1 - \frac{Y_{i1}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i}} \right). \quad (3.3)$$

A képletben a d_i azon vizsgált egyedek számát jelöli, akiknél bekövetkezett az esemény t_i időpontban. A d_{i1} pedig csak az első csoportban vizsgált ugyanilyen tulajdonságú egyedek száma. Ezután már csak a standard normális eloszlás eloszlásfüggvényével kell összevetni ezt a bizonyos z számot, ebből megkapjuk p értékét, ami megegyezik $2 \cdot (1 - \Phi(z))$ -vel. És p segítségével pedig tudunk dönteni a nullhipotézisről, attól függően, hogy $p \geq 0,05$ vagy $p < 0,05$.

3.1.1. Példa. Vegyük megint a mellrákos adatsort (forrás: [3]), és vizsgáljuk a fenti kétmintás módszerrel a hormonkezelteket és a nem hormonkezelteket. Legyen H_0 az a hipotézis, hogy a hormonkezelés nem befolyásolja a rák kiújulását és a túlélést (az esemény ebben a példában a betegség kiújulása vagy elhalálozás, illetve a páciensek megfigyelései a betegség felfedezésekor kezdődnek).

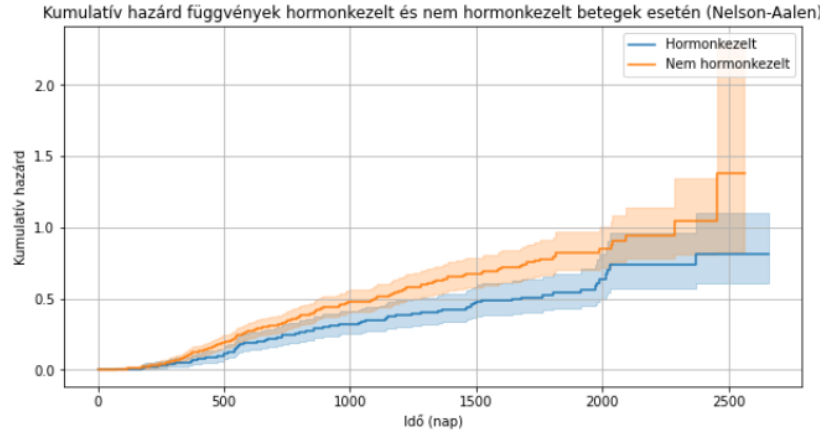
Mielőtt elvégeznénk a kétmintás próba lépéseit, készítettem kettő ábrát, melyek a fent taglalt becslők segítségével jöttek létre. Kaplan-Meier becslővel (3.2) a következő túlélési görbéket kaptam:



3.1. ábra. Túlélési görbék Kaplan-Meier becslővel

Ha rápillantunk az ábrára, akkor le tudjuk olvasni, hogy akiket hormonálisan kezeltek (kék színű függvény), azok nagyobb valószínűséggel gyógyultak meg. Ez alapján az a sejtésünk lehet, hogy a H_0 el lesz utasítva, azaz a hormonkezelés számít a túlélésben.

Ezután nézzük meg a Nelson-Aalen becslővel (3.1) készült kumulatív kockázati függvényeket is:



3.2. ábra. Kumulatív kumulatív függvények Nelson-Aalen becslővel

Ennél az ábránál azt tudjuk észrevenni, hogy a nem hormonkezeltek nagyobb eséllyel halnak meg vagy újul ki náluk a daganat. Tehát itt is arra a következtetésre tudunk jutni, hogy számít a hormonális kezelés a gyógyulásban, azaz H_0 el lesz utasítva.

Most már megvan a sejtésünk, úgyhogy kiszámolhatjuk a statisztikánkat, amivel igazolni tudjuk a várt eredményünket. Ehhez szükségünk lesz egy súlyfüggvényre, ami először legyen azonosan 1, ami a log-rank statisztikát fogja adni. Erre van külön beépített függvény a pythonban, amivel pár sor segítségével is megkapjuk a p-értéket, de most inkább végigvezetem a számolást lépésről lépésre. Legelőször meg kell határozni τ értékét, ami jelen esetben 2563 (python segítségével lett ez az eredmény). Majd ezután a D -t is kiszámoljuk: 2456 lett (szintén pythonnal jött ki ez az eredmény). Ezek után elkészítjük a segédtáblázatot, aminek D darab sora van:

idő	V_{i1}	d_{i1}	V_{i2}	d_{i2}	V_i	d_i	$V_{i1} \cdot (d_i / V_i)$	$d_{i1} - V_{i1} \cdot (d_i / V_i)$	$(V_{i1} / V_i) \cdot (1 - (V_{i1} / V_i)) \cdot ((V_i - d_i) / (V_i - 1)) \cdot d_i$
1	244.0	0.0	440.0	0.0	684.0	0.0	0.0	0.0	0.0
2	244.0	0.0	440.0	0.0	684.0	0.0	0.0	0.0	0.0
3	244.0	0.0	440.0	0.0	684.0	0.0	0.0	0.0	0.0
4	244.0	0.0	440.0	0.0	684.0	0.0	0.0	0.0	0.0
5	244.0	0.0	440.0	0.0	684.0	0.0	0.0	0.0	0.0
6	244.0	0.0	440.0	0.0	684.0	0.0	0.0	0.0	0.0
7	244.0	0.0	440.0	0.0	684.0	0.0	0.0	0.0	0.0
8	244.0	0.0	440.0	0.0	684.0	0.0	0.0	0.0	0.0
9	244.0	0.0	439.0	0.0	683.0	0.0	0.0	0.0	0.0
10	244.0	0.0	439.0	0.0	683.0	0.0	0.0	0.0	0.0

A táblázat a következő információkat tartalmazza:

- idő (1-től D -ig, ami most 2456, mert ekkor volt az utolsó esemény időpontja)
- Y_{i1} : azon hormonkezelte betegek száma, akiket még figyelnek t_i időpontban
- d_{i1} : azon hormonkezelte betegek száma, akiknek t_i időpontban kiújult a rákjuk vagy meghaltak
- Y_{i2} : azon nem hormonkezelte betegek száma, akiket még figyelnek t_i időpontban
- d_{i2} : azon nem hormonkezelte betegek száma, akiknek t_i időpontban kiújult a rákjuk vagy meghaltak
- Y_i : összes vizsgált beteg a t_i időpontban
- d_i : összes olyan beteg száma, akinek t_i időpontban kiújult a rákja vagy meghalt
- további oszlopok segédszámításokat tartalmaznak a statisztikánkhoz az utóbbi adatokat felhasználva

Miután megvan ez a nagyon hasznos táblázat, meg tudjuk határozni z értékét (3.3): az utolsó előtti oszlop összege fogja adni a z számlálóját és az utolsó oszlop összege négyzetgyök alatt pedig a z nevezőjét, mert

$$z = \frac{\sum_{i=1}^{2456} W(t_i) \left(d_{i1} - Y_{i1} \frac{d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^{2456} W(t_i)^2 \frac{Y_{i1}}{Y_i} \left(1 - \frac{Y_{i1}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i}}. \quad (3.4)$$

Ezt szintén python segítségével számoltam ki, és $z = -2,790$ lett az eredmény. Most jöhet a p -érték meghatározása, amit úgy kapunk meg, hogy vesszük a $Z > z$ valószínűség kétszeresét (ha z nem negatív, különben a standard normális eloszlás sűrűségfüggvény szimmetriája miatt a $Z \leq z$ valószínűség kétszeresét kell kiszámolni), ahol Z standard normális eloszlású valószínűségi változó. Tehát $p = 2 \cdot \mathbb{P}(Z \leq z) = 2 \cdot \Phi(z) = 2 \cdot (0,0016) = 0,0032$, ahol

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{s^2}{2}} ds$$

a standard normális eloszlás eloszlásfüggvénye. Ahhoz, hogy el tudjuk dönteni a hipotézisünket p -t össze kell még vetni 0,05-tel. És mivel kisebb a p , ezért elutasítjuk H_0 -t, azaz a sejtésünk beigazolódott - tényleg van különbség a hormonkezeltek és a nem hormonkezeltek túlélése között, nem felesleges a terápia.

3.1.2. Példa. Legyen teljesen ugyanaz a feladat, mint a fenti példában, csak a súlyfüggvényt változtassuk meg a Gehan-félére: $W(t_i) = Y_i$ (az előző példában $W(t_i) = 1$ volt). Ehhez egy újabb segédtáblázatot kell készíteni, mert a súlyfüggvény szerepel z számlálójában és a nevezőjében is. Nem sokban különbözik az új táblázat az előzőtől (3.1.1) - csak az utolsó 2 oszlopban az eredeti kifejezéseket megszoroztam Y_i -vel. Ezek után z megint az utolsó 2 oszlop segítségével megadható, így $z = -2,803$ lett az eredmény, ami nem meglepő módon elég közel van az előző példa z értékéhez ($-2,790$). A p -érték számítása ugyanúgy megy, mint korábban, és 0,005 jön ki rá. Ennek következtében ezzel a súlyfüggvénnyel is ugyanazt a végeredményt kapjuk, miszerint H_0 -t elutasítjuk.

3.1.3. Példa. Megint a mellrákos adatsort vizsgáljuk (forrás: [3]). De most a gyógyulást a hormonkezelés helyett a kor függvényében nézzük. Ehhez a 2 csoport az 50 év alattiak és az 50+-osok. Tehát H_0 legyen az a nullhipotézis, hogy a kor nem befolyásolja a túlélést. A fenti eljárást alkalmazva $p = 0,159$ lett, ami már nagyobb, mint 0,05, tehát H_0 -t elfogadjuk 5%-os szignifikanciaszint mellett. Így nincs törvényszerű eltérés a fiatalok és az idősek gyógyulásában.

3.2. Három vagy több mintás teszt

A kétmintás változathoz hasonlóan működik ez a teszt. Itt $H_0 : h_1(t) = h_2(t) = \dots = h_K(t) \forall t \leq \tau$ esetén, és $H_A : h_j(t)$ értékek közül legalább az egyik különbözik bizonyos t -re. Ugyanazok a súlyfüggvények használhatóak itt is, mint a kétmintás esetnél. Azonban most több z értékünk is lesz, mert az összes lehetséges csoportpárra ki kell számolni az eredményt ezzel a képlettel:

$$z_j(\tau) = \sum_{i=1}^D W(t_i) \left[d_{ij} - Y_{ij} \frac{d_i}{Y_i} \right], \text{ ahol } j = 1, \dots, K. \quad (3.5)$$

Ez a bonyolult kifejezés a megfigyelt halálesetek száma és a j -edik mintában a halálesetek várható száma közötti súlyozott különbség összegét adja. Ha az összes $z_j(\tau)$ közel van a nullához, akkor kevés bizonyíték van a nullhipotézis hamisságára, hiszen ekkor minden csoportban a csoport számával arányosan következik be a halálozás (d_{ij} -vel szokták jelölni). Ezzel szemben, ha létezik j , hogy $z_j(\tau)$ értéke távol van a 0-tól, akkor el fogjuk utasítani H_0 -t, mivel ez bizonyíték arra, hogy a populáció hazardráta eltér a nullhipotézisben feltettektől. Ha sikeresen meghatároztuk a $z_j(\tau)$ -kat, akkor a kovarianciákat, illetve varianciákat is ki kell számolni úgy, hogy

$$\sigma_{jg} = - \sum_{i=1}^D W(t_i)^2 \frac{Y_{ij}}{Y_i} \frac{Y_{ig}}{Y_i} \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i, \quad j \neq g \quad (3.6)$$

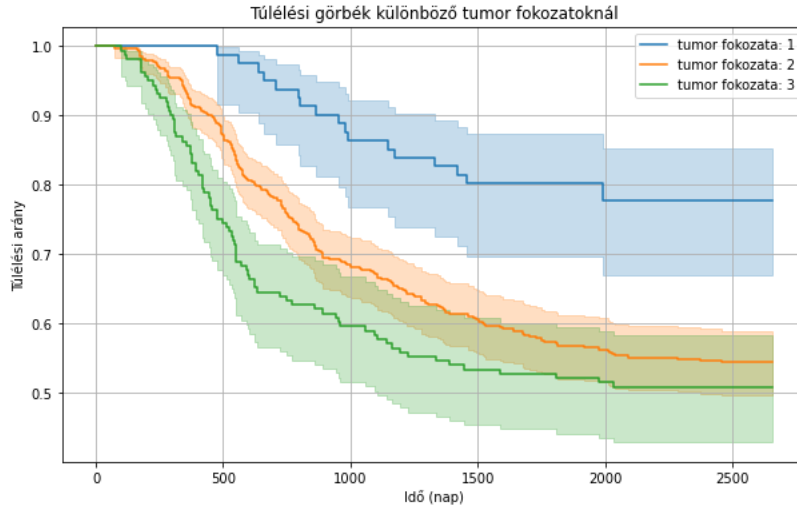
$$\sigma_{jj} = - \sum_{i=1}^D W(t_i)^2 \frac{Y_{ij}}{Y_i} \left(1 - \frac{Y_{ij}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i, \quad j = 1, \dots, K. \quad (3.7)$$

Ebből már el tudjuk készíteni a $(K - 1) \times (K - 1)$ -es variancia-kovariancia mátrixot úgy, hogy tetszőleges $K - 1$ darab z_j -t kiválasztunk, és ezeknek számoljuk ki a különböző σ értékeiket. Ezt a mátrixot jelöljük Σ^{-1} -zel. Ezután a χ^2 -et határozzuk meg az alábbi szorzat segítségével:

$$\chi^2 = (z_1(\tau), \dots, z_{K-1}(\tau)) \Sigma^{-1} (z_1(\tau), \dots, z_{K-1}(\tau))^T. \quad (3.8)$$

H_0 esetén a próbastatisztikánk χ^2 eloszlású lesz $K - 1$ szabadsági fokkal.

3.2.1. Példa. Vegyük megint a mellrákos adatsort (forrás: [3]). Viszont most a tumor fokozatai alapján készítjük el a tesztet. Ehhez is készítettem túlélési görbéket a Kaplan-Meier becslővel (3.2):



3.3. ábra. Túlélési görbék Kaplan-Meier becslővel

Az ábráról leolvasható, hogy minél nagyobb fokozatú a tumor, annál kevésbé valószínű a rákból való meggyógyulás. Ezt a sejtést most számokkal is alátámasztjuk. A z_j -k kiszámításához megint készítettem egy táblázatot, de ennek sajnos már sokkal több oszlopa van, mint az eddigieknek, mert az összes lehetséges csoportpárosítást végig kell nézni. Ezután a $z_j(\tau)$ képlete segítségével (3.5) meg tudjuk határozni a z_j -ket (pythonnal számoltam ki): $z_1 = -24,112$, $z_2 = 4,900$ és $z_3 = 19,212$. Majd a variancia-kovariancia mátrix elemeit meghatározva létrehozuk a $(K-1) \times (K-1) = (3-1) \times (3-1) = 2 \times 2$ -es részmátrixot. Az eredeti így néz ki:

$$\begin{bmatrix} 36.08284741 & -27.75198585 & -8.33086156 \\ -27.75198585 & 67.06303652 & -39.31105067 \\ -8.33086156 & -39.31105067 & 47.64191223 \end{bmatrix}$$

A z_1 és z_2 által meghatározott kovarianciák részmátrixa (Σ^{-1}) pedig a következő:

$$\begin{bmatrix} 36,08284741 & -27,75198585 \\ -27,75198585 & 67,06303652 \end{bmatrix}$$

Most már csak χ^2 -et kell meghatározni (3.8), ami nem más, mint $(z_1(\tau), z_2(\tau))\Sigma^{-1}(z_1(\tau), z_2(\tau))^T = 20,184$. A próbastatisztikát 5%-os szignifikanciaszint mellett nézzük 2 szabadsági fokkal, így a kritikus érték 5,991 lesz. Ennél jóval nagyobb a mi értékünk, tehát a nullhipotézist elutasítjuk, azaz

tényleg van különbség azon betegek gyógyulása között, akiknek más fokozatú a tumorjuk.

3.3. Trend tesztelése

Most egy kicsit másfajta statisztikát vizsgálunk, mert itt már a rendezett alternatívákat is ki tudjuk mutatni (azaz nem csak azt nézzük meg, hogy van-e szignifikáns különbség a különböző csoportokban a túlélés terén, hanem azt is, hogy melyik csoportban van a legnagyobb esély, illetve a legkevesebb esély a túlélésre rögzített időpontokban). Ezért $H_0 : h_1(t) = h_2(t) = \dots = h_K(t) \forall t \leq \tau$, és $H_A : h_1(t) \leq h_2(t) \leq \dots \leq h_K(t) \forall t \leq \tau$ legalább egy szigorú egyenlőtlenséggel. A teszthez a (3.5) által megadott $z_j(\tau)$ -kra van szükség. A korábbi súlyfüggvények ugyanúgy itt is használhatóak, amiknek az a lényege, hogy súlyozzák azokat az időpontokat, amikor az egyes mintákban megfigyelt és várható halálesetek számát összehasonlítjuk. Szintén az előző 3 mintás statisztikában alkalmazott kovariancia mátrixra itt is szükségünk lesz. Ami újdonság, hogy ki kell választani egy szigorúan növekvő pontsorozatot: $a_1 < a_2 < \dots < a_K$, ami a populációk numerikus jellemzője lesz. Bármilyen pontszámkészlet használható, ami az utóbbi feltételt teljesíti (a legtöbb esetben $a_j = j$ -t választják). A próbastatisztika a következő lesz (a σ_{jg} -t 3.6-ban definiáltuk):

$$z = \frac{\sum_{j=1}^K a_j z_j(\tau)}{\sqrt{\sum_{j=1}^K \sum_{g=1}^K a_j a_g \sigma_{jg}}}. \quad (3.9)$$

Ha a nullhipotézis igaz, akkor ennek a statisztikának standard normálisnak kell lennie.

3.3.1. Példa. *Mellrákos adatsorban (forrás: [3]) most azt vizsgáljuk, hogy a tumor fokozat növekedésével egyre kevesebb-e az esély a meggyógyulásra. A 3.2.1-es korábbi példában kiszámoltuk a kovariancia mátrixot és a z -ket is. A pontsorozat legyen a következő: $a_1 = 1$, $a_2 = 2$ és $a_3 = 3$. A fenti képlet alapján (3.9) $z = 4,324$, ebből $p = 2 \cdot \mathbb{P}(Z > z) = 2 \cdot (1 - \Phi(z)) = 1,533 \cdot 10^{-5}$. Tehát p nagyon kicsi, ami azt jelenti, hogy 5%-os szignifikanciaszint*

mellett elutasítjuk H_0 -t, azaz minél kisebb a tumor fokozata, annál nagyobb valószínűséggel gyógyul meg a beteg.

3.4. Kétlépcsős eljárás

Az előbbi statisztikai módszereknél főleg a log-rank tesztet alkalmaztuk. Azonban ezt az eljárást csak olyan esetben szabad használni, amikor a túlélési függvények nem keresztezik egymást, azaz amikor a csoportok hazard függvényei arányosak (ebben az esetben a csoportok közötti különbségek kimutatására tényleg a log-rank teszt a legerősebb módszer). Másfajta esetekben - különösen kereszteződő hazard függvényeknél - nem működik a teszt, mert ilyenkor a H_0 visszautasítási aránya jelentősen csökken. Ennek ellenére nagyon sokan továbbra is használják ezt a módszert helytelenül keresztező túlélési görbék esetén is (Kristiansen kutató 175 tanulmányt tekintett át, és az esetek 70%-ában rosszkor használták a log-rank tesztet a [6] forrás alapján).

Nem lehet azt mondani, hogy annyira kevésszer esik meg a kereszteződés, hogy nem fontos ezzel külön foglalkozni, mert az onkológiában is számtalanszor előfordul (például amikor egy kezelés rövid távon előnyöket biztosíthat, de hosszú távon már nem nyújt előnyöket - sugárkezelés, illetve kemoterápia is általában ide tartozik). Persze olyan is előfordulhat az orvosi kezeléseknél, hogy rövid távon növelik a halálozás kockázatát, viszont hosszú távon előnyökkel járnak (ilyen például egy műtét is). Ennek következtében ezen esetek kezelésére számos módszert kifejlesztettek (például a Neyman-féle sima teszt, Kolmogorov-Smirnov-teszt vagy a kétlépcsős eljárás is ilyen). A továbbiakban a kétlépcsős módszer részletes bemutatása a cél (nagyraért [8] forrás alapján).

A kétlépcsős eljárás azért egy „adúasz” módszer, mert nemcsak keresztező görbék esetén alkalmazható, hanem különböző, de nem keresztező görbék eseténél is. Az eljárás nevéből adódóan két szakaszból áll a próbastatisztika. Az első részben egy „hagyományos” eljárást alkalmazunk (például a log-rank tesztet vagy a Peto-Peto-tesztet ...), a második részben pedig az átfedések kimutatására szolgáló eljárást hajtjuk végre.

Legyen h_0 és h_1 a két vizsgált csoportba tartozó alanyok túlélési idejének hazard függvénye, és legyen a vizsgált időtartomány $[0, \tau]$. Ezután a következő hipotéziseket vizsgáljuk: $H_0 : h_0(t) = h_1(t) \forall t \in [0, \tau]$ és

$H_A : h_0(t) \neq h_1(t)$ (ez az alternatív hipotézis magába foglalja az összes lehetséges esetet - keresztező és párhuzamosan futó görbék is).

Úgy fogjuk elvégezni az eljárás „lépcsőt”, hogy ha az első részben elutasításra kerül a nullhipotézis, akkor nem is megyünk tovább a második részhez, hanem arra a következtetésre jutunk, hogy a két kártya függvény szignifikánsan különbözik egymástól. Ellenkező esetben a második szakaszba lépünk, és itt meg tudjuk különböztetni azokat az eseteket, amikor a két kártya függvény megegyezik, illetve amikor keresztezik egymást. (Ez a végrehajtási sorrend nem kőbe vésett, a sorrend megváltoztatásával is működik az eljárás.)

3.4.1. Számolás menete

Legyen két csoport ($j = 1, 2$), és legyen n_j a j . csoportban lévő alanyok száma, így $n = n_1 + n_2$. Tegyük fel, hogy $t_1 < t_2 < \dots < t_D$ a különböző rendezett eseményidőpontok halmaza az összevont mintában. Az $i = 1, \dots, D$ értékek esetében legyen d_{ij} továbbra is a j . csoportban a t_i időpontban bekövetkezett események száma ($d_i = d_{i1} + d_{i2}$). Emellett Y_{ij} a korábbi definiáláshoz hasonlóan jelölje a j . csoportban a t_i -ben megfigyelt egyedek számát ($Y_i = Y_{i1} + Y_{i2}$).

Ekkor az első „lépcső”-höz használt log-rank teszt próbastatisztikája a következő (ez ugyanaz, mint (3.3)-as képlet):

$$U = \frac{\sum_{i=1}^D W_{i1}(t_i) \left(d_{i1} - Y_{i1} \frac{d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^D W_{i1}(t_i)^2 \frac{Y_{i1}}{Y_i} \left(1 - \frac{Y_{i1}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i}}, \quad (3.10)$$

ahol $W_{i1}(t_i) = 1 \forall t_i$ -re.

A második „lépcső” célja a kereszteződési görbék kezelése. Olyan eljárás levezetése következik, amire teljesül, hogy aszimptotikusan független az első „lépcsőtől”, azaz a log-rank tesztől (ennek a bizonyítása a 3.4.1 tételnél van). A fenti jelöléseket felhasználva legyen $k = 1, \dots, n_j$, ekkor T_{kj} a j . csoport k . egyedének eseményideje F_j eloszlásfüggvény mellett, illetve C_{kj} a cenzorálás ideje G_j eloszlásfüggvény mellett. Továbbá legyen $S_j(s) = 1 - F_j(s)$ és $L_j(s) = 1 - G_j(s)$. Kell választani egy 0-nál nagyobb és 0,5-nél kisebb ε -t, ekkor $D_\varepsilon = [D * \varepsilon]$ (az eredmények meglehetősen stabilak lesznek, ha ε -t úgy

választjuk, hogy $D_\varepsilon \geq 5$). Bármely $r \in [\varepsilon, 1 - \varepsilon]$ esetén legyen $m = [D * r]$, így fennáll a következő összefüggés: $D_\varepsilon \leq m \leq D - D_\varepsilon$. A próbastatisztikánk új súlyozása legyen a következő:

$$W_{i2}^m = \begin{cases} -1, & \text{ha } i = 1, 2, \dots, m \\ c_m, & \text{különben,} \end{cases} \quad (3.11)$$

ahol c_m pozitív mennyiség. Az alábbi becslés segítségével tudunk c_m -re közelítő értéket adni:

$$\hat{c}_m = \frac{\sum_{i=1}^m \frac{\hat{L}_1(t_i) \hat{L}_2(t_i)}{\frac{n_1}{n} \hat{L}_1(t_i) + \frac{n_2}{n} \hat{L}_2(t_i)} \cdot \Delta \hat{S}(t_i)}{\sum_{i=m+1}^D \frac{\hat{L}_1(t_i) \hat{L}_2(t_i)}{\frac{n_1}{n} \hat{L}_1(t_i) + \frac{n_2}{n} \hat{L}_2(t_i)} \cdot \Delta \hat{S}(t_i)}, \quad (3.12)$$

ahol $\hat{L}_j$ és $\hat{S}$ a cenzorálás és az eseményidők túlélési függvényeinek a Kaplan-Meier becslései (kiszámításukhoz a képlet itt megtalálható: (3.2)). Miután sikerült meghatározzuk a fenti értékeket, a próbastatisztikánkat is ki tudjuk számolni az alábbi módon:

$$V_m = \frac{\sum_{i=1}^D W_{i2}^m(t_i) \left(d_{i1} - Y_{i1} \frac{d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^D (W_{i2}^m(t_i))^2 \frac{Y_{i1}}{Y_i} \left(1 - \frac{Y_{i1}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i}}, \quad (3.13)$$

és

$$V = \sup_{D \leq m \leq D - D_\varepsilon} V_m. \quad (3.14)$$

Az U és a V próbastatisztikát azért választottuk a fenti módon (3.10 és 3.14), mert így az alábbi tétel igaz:

3.4.1. Tétel. *(A tétel és a bizonyítás is a [8] forrás alapján készült.) Tegyük fel, hogy $j = 1, 2$ esetén az eseményidő eloszlásfüggvénye F_j és a cenzúraidőnek pedig G_j . A két csoport megfigyelései függetlenek egymástól, és a cenzúrázási időpontok is függetlenek a bekövetkező események időpontjaitól. Ekkor U és V aszimptotikusan független, feltételezve, hogy $F_1 \equiv F_2$.*

3.4.2. Lemma.

$$\frac{\sigma^2(\underline{aw}_1 + \underline{bw}_2^r)}{\sigma^2(\underline{w}_1)} \xrightarrow{P} a^2 + b^2 \text{ ha } n \rightarrow \infty, \quad (3.15)$$

(A fenti lemma bizonyítása megtalálható a [8] forrásban.)

Bizonyítás. A tétel bizonyításához következő jelölésekre lesz szükség:

- $S_j(s) = 1 - F_j(s)$
- $L_j(s) = 1 - G_j(s)$
- $X_{kj} = \min(T_{kj}, C_{kj})$, ahol T_{kj} a j . csoport k . egyedének eseményideje, C_{kj} pedig a cenzorálás ideje ugyanazon egyednél
- $\delta_{kj} = \mathbb{1}_{\{T_{kj} < C_{kj}\}}$ (akkor 1 az érték, ha hamarabb következik be az esemény, mint a cenzúra időpontja)
- $\pi_j(s) = \mathbb{P}(X_{kj} > s) = L_j(s)S_j(s)$, ez annak a valószínűsége, hogy az adott egyednél s időpontnál tovább tart a megfigyelés

Legyen $\underline{w} = (w_1, \dots, w_D)^T$ (ez a súlyokat tartalmazó vektor). Legyen

$$Z(\underline{w}) = h \sum_{i=1}^D w_i \left(d_{i1} - Y_{i1} \frac{d_i}{Y_i} \right)$$

(ez a h nélkül az U számlálója) és

$$\hat{\sigma}^2(\underline{w}) = h^2 \sum_{i=1}^D w_i^2 \frac{Y_{i1}}{Y_i} \frac{Y_{i2}}{Y_i} \frac{Y_i - d_i}{Y_i - 1} d_i$$

(ez pedig h^2 nélkül az U nevezője gyök alatt), ahol $h = \sqrt{\frac{n}{n_1 \cdot n_2}}$. Továbbá

legyen $\overline{Y}_j(s) = \sum_{k=1}^{n_j} \mathbb{1}_{\{X_{kj} \geq s\}}$ (ez a kifejezés azt jelenti, hogy j . csoportban

hány egyedet vizsgáltak legalább s időpontig) és $\overline{N}_j(s) = \sum_{k=1}^{n_j} \mathbb{1}_{\{X_{kj} \leq s, \delta_{kj}=1\}}$

(ez pedig azt írja le, hogy a j . csoportban hány esemény következett be s -ig).

Legyen $\hat{S}(s)$ az $S(s)$ túlélési függvény Kaplan-Meier becslése (3.2), továbbá $W(s)$ legyen az $\hat{S}(s-)$ megjósolható korlátos függvénye, melyre teljesül, hogy $(W(t_1), \dots, W(t_D))^T = \underline{w}$ (azaz $\underline{w}$ -t kiterjesztjük függvénnyé az egész intervallumon úgy, hogy $W(s)$ csak az $\hat{S}$ s -beli baloldali határértékétől függ).

Ekkor

$$Z(\underline{w}) = h \int_0^u W(s) \frac{\overline{Y}_1(s) \overline{Y}_2(s)}{\overline{Y}_1(s) + \overline{Y}_2(s)} \left\{ \frac{d\overline{N}_1(s)}{\overline{Y}_1(s)} - \frac{d\overline{N}_2(s)}{\overline{Y}_2(s)} \right\}, \quad (3.16)$$

ahol $u = \inf\{s : \min(\pi_1(s), \pi_2(s)) = 0\}$.

Legyen

$$K_W(s) = hW(s) \frac{\overline{Y}_1(s) \overline{Y}_2(s)}{\overline{Y}_1(s) + \overline{Y}_2(s)}.$$

$\frac{\overline{Y}_1(s)}{n_1} \pi_1(s)$ konzisztens becslője (azaz sztochasztikusan tart hozzá $n \rightarrow \infty$ esetén), $\frac{\overline{Y}_2(s)}{n_2} \pi_2(s)$ konzisztens becslője és $\frac{\overline{Y}_1(s) + \overline{Y}_2(s)}{n}$ pedig $p_1 \pi_1(s) + p_2 \pi_2(s)$ konzisztens becslője, ahol $p_j = \lim_{n \rightarrow \infty} \frac{n_j}{n}$. Az utóbbi azért igaz, mert

$$\frac{\overline{Y}_1(s) + \overline{Y}_2(s)}{n} = \frac{\overline{Y}_1}{n_1} \cdot \frac{n_1}{n} + \frac{\overline{Y}_2}{n_2} \cdot \frac{n_2}{n} \xrightarrow{P} \pi_1(s)p_1 + \pi_2(s)p_2.$$

Ebből következik, hogy

$$\frac{K_W^2(s)}{\overline{Y}_j(s)} = \frac{\frac{n}{n_1 n_2} \cdot W^2(s) \cdot \frac{\overline{Y}_1^2(s) \overline{Y}_2^2(s)}{(\overline{Y}_1(s) + \overline{Y}_2(s))^2}}{\overline{Y}_j(s)} \quad (3.17)$$

ugyanoda tart sztochasztikusan, mint

$$\frac{W^2(s) \cdot \frac{\pi_1^2(s) \pi_2^2(s)}{(p_1 \pi_1(s) + p_2 \pi_2(s))^2} \cdot \frac{n_1 n_2}{n}}{\overline{Y}_j(s)}. \quad (3.18)$$

Másrészt

$$\frac{W^2(s) \cdot \frac{\pi_1^2(s) \pi_2^2(s)}{(p_1 \pi_1(s) + p_2 \pi_2(s))^2} \cdot \frac{n_1 n_2}{n}}{\overline{Y}_j(s)} \quad (3.19)$$

$$\xrightarrow{P} W^2(s) \cdot \frac{p_1 p_2 \pi_1^2(s) \pi_2^2(s)}{p_j \pi_j(s) (p_1 \pi_1(s) + p_2 \pi_2(s))^2}, \quad \text{ha } n \rightarrow \infty \text{ és } j = 1, 2.$$

Fleming és Harrington (1991) 7.2.1 tételének (1)-(3) regularitási feltételei teljesülnek ([9] forrásban található), így

$$\frac{Z(\underline{w})}{\sigma(\underline{w})} \xrightarrow{d} N(0, 1). \quad (3.20)$$

és

$$\begin{aligned} \sigma^2(\underline{w}) &= \int_0^u W^2(s) \frac{\pi_1(s)\pi_2(s)}{p_1\pi_1(s) + p_2\pi_2(s)} \cdot (1 - \Delta\Lambda(s)) d\Lambda(s) \\ &= \int_0^u W^2(s) \frac{\pi_1(s)\pi_2(s)}{p_1\pi_1(s) + p_2\pi_2(s)} \cdot \frac{1}{S(s)} dF(s) \\ &= \int_0^u W^2(s) \frac{L_1(s)L_2(s)}{p_1L_1(s) + p_2L_2(s)} dF(s), \end{aligned}$$

ahol $\Lambda(s)$ az eseményidő közös kumulatív hazárd függvénye H_0 esetén, ami folytonos.

Minden $r \in [\varepsilon, 1 - \varepsilon]$ -ra legyen $m = [D \cdot r]$ és $W_2^r(s) = -\mathbb{1}_{\{s < t_m\}} + \hat{c}_m \cdot \mathbb{1}_{\{s \geq t_m\}}$. Ekkor $W_2^r(s)$ az $\hat{S}(s-)$ jósolható függvénye.

Minden $\varepsilon \leq r \leq 1 - \varepsilon$ -ra $\sigma^2(\underline{w}) = \int_0^u W^2(s) \frac{L_1(s)L_2(s)}{p_1L_1(s) + p_2L_2(s)} dF(s) < \infty$, mert $\hat{c}_m < \infty$.

Létezik egy konstans $w_0 \in (0, \infty)$: ha $W_1(s) \equiv w_0$, akkor

$$\sigma^2(\underline{w}_1) = \sigma^2(\underline{w}_2^r), \quad (3.21)$$

ahol $\underline{w}_1 = (W_1(t_1), \dots, W_1(t_D))^T = w_0 \cdot \mathbb{1}_D$ és $\underline{w}_2^r = (W_2^r(t_1), \dots, W_2^r(t_D))^T$. Hiszen ha $\underline{w}_1 = (W_1(t_1), \dots, W_1(t_D))^T = w_0 \cdot \mathbb{1}_D$ konstans, akkor a $\sigma^2(\underline{w}_1)$ integrál ennek a négyzetétől lineárisan függ, ezért w_0 -t változtatva találunk megfelelő értéket.

Legyen $U^* = \frac{Z(\underline{w}_1)}{\sigma(\underline{w}_1)}$ és $V_r^* = \frac{Z(\underline{w}_2^r)}{\sigma(\underline{w}_2^r)}$.

A következőkben belátjuk, hogy

$$\begin{pmatrix} U^* \\ V_r^* \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, I_2\right), \text{ ha } n \rightarrow \infty. \quad (3.22)$$

Tetszőleges a, b konstansok mellett, 3.21 alapján következik, hogy

$$aU^* + bV_r^* = \frac{Z(a\underline{w}_1 + b\underline{w}_2^r)}{\sigma(\underline{w}_1)}. \quad (3.23)$$

Továbbá 3.16 és 3.20 miatt teljesül, hogy

$$\frac{Z(\underline{aw}_1 + \underline{bw}_2^r)}{\sigma(\underline{aw}_1 + \underline{bw}_2^r)} \xrightarrow{d} N(0, 1), \text{ ha } n \rightarrow \infty. \quad (3.24)$$

A 3.4.2 lemmát felhasználva és 3.23, 3.24 alapján

$$aU^* + bV_r^* \xrightarrow{d} N(0, a^2 + b^2), \text{ ha } n \rightarrow \infty. \quad (3.25)$$

Ebből pedig következik 3.22, amit szerettünk volna belátni.

Tudjuk, hogy $U = U^* \frac{\sigma(\underline{w}_1)}{\hat{\sigma}(\underline{w}_1)}$ és $V_m = V_r^* \frac{\sigma(\underline{w}_2^r)}{\hat{\sigma}(\underline{w}_2^r)}$, mivel

$$\begin{aligned} U &= \frac{\sum_{i=1}^D W_{i1}(t_i) \left(d_{i1} - Y_{i1} \frac{d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^D W_{i1}(t_i)^2 \frac{Y_{i1}}{Y_i} \left(1 - \frac{Y_{i1}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i}} = \\ &= \frac{Z(\underline{w}_1)}{\hat{\sigma}(\underline{w}_1)} = \frac{Z(\underline{w}_1)}{\sigma(\underline{w}_1)} \cdot \frac{\sigma(\underline{w}_1)}{\hat{\sigma}(\underline{w}_1)} = U^* \cdot \frac{\sigma(\underline{w}_1)}{\hat{\sigma}(\underline{w}_1)} \end{aligned}$$

(ugyanilyen meggondolással kijön a V_m is). Fleming és Harrington (1991) 7.2.1 tétele alapján ([9]) teljesül, hogy

$$\frac{\sigma(\underline{w})}{\hat{\sigma}(\underline{w})} \xrightarrow{P} 1, \text{ ha } n \rightarrow \infty, \quad (3.26)$$

$\underline{w} = \underline{w}_1$, illetve $\underline{w} = \underline{w}_2$ esetén is. Van der Vaart (1998) 18.10-es és 18.11-es tételéből ([10]) következik, hogy (U, V_m) konvergál egy 0 várható értékű és identitás kovarianciamátrixú kétváltozós normális vektorváltozóhoz. Emiatt U és V_m aszimptotikusan függetlenek egymástól (mivel $cov(U, V_m) \rightarrow 0$ és (U, V_m) együttesen normális eloszlásúak), amiből következik, hogy U és V is aszimptotikusan független. $\square$

3.4.2. Szignifikanciaszint és p-érték meghatározása

A kétlépcsős eljárás szignifikanciaszintje legyen α (általában $\alpha = 0,05$), és a két szakaszának a szignifikanciaszintje pedig legyen α_1 és α_2 . Így teljesül, hogy

$$\alpha_1 + \mathbb{P}(\text{elutasítjuk } H_0 \text{-t} | \text{első szakaszban nem utasítottuk el } H_0 \text{-t})(1 - \alpha_1) = \alpha. \quad (3.27)$$

A 3.4.1 tétel alapján tudjuk, hogy az U és a V próbastatisztika aszimptotikusan független (azaz $\lim_{n \rightarrow \infty} \mathbb{P}(U \leq s, V \leq t) = \lim_{n \rightarrow \infty} \mathbb{P}(U \leq s) \cdot \mathbb{P}(V \leq t)$ $\forall s, t$ esetén), ezért igaz az, hogy

$$\alpha_1 + \alpha_2(1 - \alpha_1) = \alpha. \quad (3.28)$$

Az alkalmazásokban α_1 és α_2 meghatározható, ha van előzetes információnk a hazárd rátákról. Ha például úgy gondoljuk, hogy a két hazárd függvény nem keresztezheti egymást, vagy megegyeznek, akkor az α_1 választható α -nak, illetve 0-nak. Ha nincs ilyen előzetes információnk, akkor egyszerűen legyen

$$\alpha_1 = \alpha_2 = 1 - \sqrt{1 - \alpha}. \quad (3.29)$$

Miután meghatároztuk az α_1 -et és α_2 -t, a kritikus értékeket kell még figyelembe venni. Az első részhez az aszimptotikus standard normális U eloszlás tartozik H_0 alatt. A második szakasznál már kicsit bonyolultabb a helyzet, mert ott a bootstrap módszert kell alkalmazni (az eredeti mintából veszünk többször visszatevéssel mintákat, és azokra számoljuk ki a V próbastatisztikát).

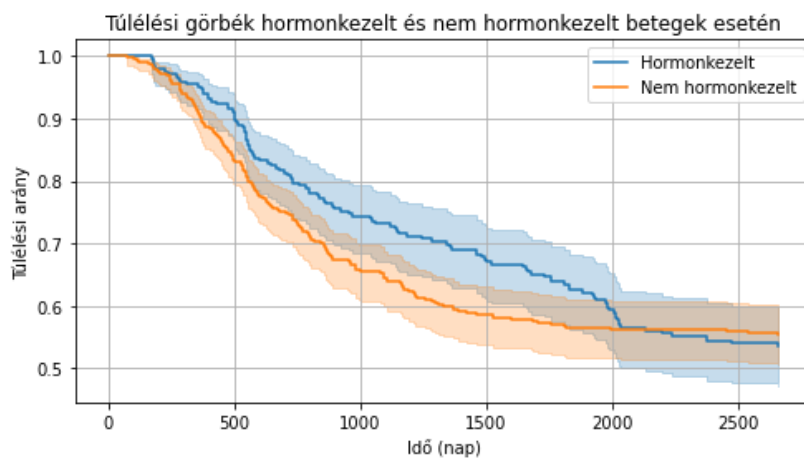
A p -érték többféleképpen meghatározható. Mivel a mi esetünkben a két különálló szakaszban végzett vizsgálatok aszimptotikusan függetlenek H_0 alatt, így az alábbi definiálás kényelmes:

$$p\text{-érték} = \begin{cases} p_1, & \text{ha } p_1 \leq \alpha_1 \\ \alpha_1 + p_2(1 - \alpha_1), & \text{különben,} \end{cases} \quad (3.30)$$

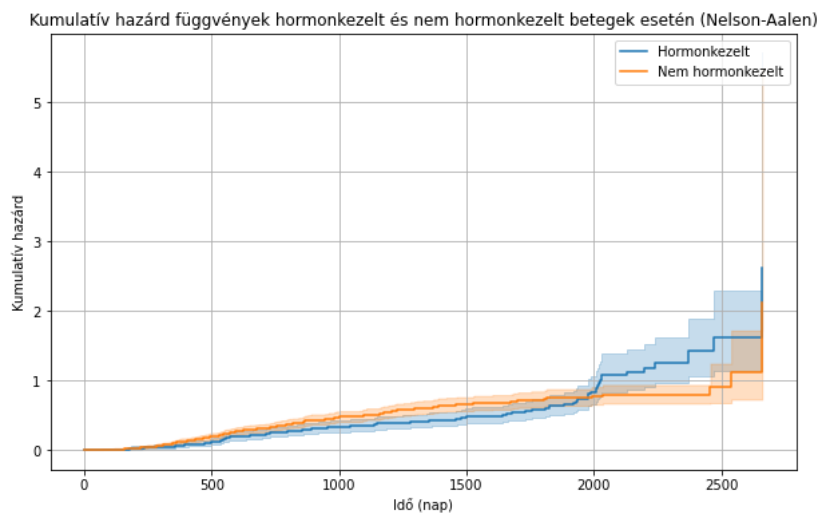
ahol p_1 és p_2 a két „lépcső” p -értéke. A fenti meghatározás alapján elutasítjuk H_0 -t, ha az első szakaszban a nullhipotézist elutasítjuk (mert ebben az esetben teljesül, hogy $p_1 \leq \alpha_1$), vagy ha az első részben elfogadtuk, de a másodikban már elutasítottuk H_0 -t (azaz $p_1 > \alpha_1$ és $p_2 \leq \alpha_2$). A 3.28 és az előbbi egyenlőtlenség alapján teljesül, hogy $\alpha_1 + p_2(1 - \alpha_1) \leq \alpha$, ami kell ahhoz, hogy H_0 -t el tudjuk utasítani. Tehát tényleg jól van definiálva a p -érték.

FIGYELEM!!! A következőben egy olyan példa lesz, ami nem valós adatokból van (a korábbi mellrákos adathalmazban átírtam a számokat, hogy keresztező túlélési görbékkel lehessen számolni), így a kialakuló eredményekből nem szabad levonni általános következtetéseket a mellrákkal kapcsolatban.

3.4.3. Példa. *Mint ahogy több korábbi példában, most is a hormonkezelés alapján vizsgáljuk meg a mellrákos betegek túlélését (de nem ugyanaz az adathalmaz). Először a Kaplan-Meier és a Nelson-Aalen becslés segítségével leellenőrizzük, hogy tényleg keresztezik-e a görbék egymást.*



3.4. ábra. Túlélési görbék Kaplan-Meier becslővel mesterségesen módosított adatokból



3.5. ábra. Kumulatív kártevő függvények Nelson-Aalen becslővel mesterségesen módosított adatokból

Az ábrák nagyon jól alátámasztják, hogy a túlélési görbéknel, illetve hazard függvényeknél van metszés.

A próbastatisztikához meg kell határozni a szignifikanciaszinteket. Legyen $\alpha = 0,05$, illetve $\alpha_1 = \alpha_2 = 1 - \sqrt{1 - \alpha} = 0,0253$.

Ezután jöhet az első „lépcső”, aminél a log-rank statisztikát használjuk. p_1 -re 0,40 lett az eredmény, ami nagyobb, mint 0,0253, így nem tudjuk elutasítani a nullhipotézist. Ennek következtében el kell végezzük a második szakaszt is.

A második „lépcső”-höz szükségünk van a D értékre (legkésőbb bekövetkezett esemény időpontja - ez most 2659-cel egyezik meg), illetve egy ε -ra, ami most legyen 0,05. Így $D_\varepsilon = [D \cdot \varepsilon] = 132$. m -et úgy választjuk meg, hogy $D_\varepsilon \leq m \leq D - D_\varepsilon$ teljesüljön, így például $m = 1329$ jó. Majd később az összes lehetséges m -re fogunk szuprérumot venni, hogy megkapjuk a próbastatisztikát.

Ahhoz, hogy az új súlyfüggvényt meg tudjuk határozni, még ki kell számolni, hogy hány pácienszt kezeltek hormonálisan (n_1 -gyel jelöljük a továbbiakban), illetve anélkül (ez pedig legyen n_2). Ezenkívül még meg kell becsülni $\hat{L}_j$ -ket és $\hat{S}$ -et a Kaplan-Meier módszer segítségével. Pythont használva kijött, hogy $n_1 = 246$ és $n_2 = 440$, illetve az alábbi táblázatban minden időpontra elkészültek a becslések:

idő	Y_i	d_i	1-(d_i/Y_i)	S(t)	Delta	S(t)	L1(t)	L2(t)
1	310.0	0.0	1.0	1.0	0.0	1.0	1.000000	1.000000
2	310.0	0.0	1.0	1.0	0.0	1.0	1.000000	1.000000
3	310.0	0.0	1.0	1.0	0.0	1.0	1.000000	1.000000
4	310.0	0.0	1.0	1.0	0.0	1.0	1.000000	1.000000
5	310.0	0.0	1.0	1.0	0.0	1.0	1.000000	1.000000
6	310.0	0.0	1.0	1.0	0.0	1.0	1.000000	1.000000
7	310.0	0.0	1.0	1.0	0.0	1.0	1.000000	1.000000
8	310.0	0.0	1.0	1.0	0.0	1.0	0.995902	0.995902
9	310.0	0.0	1.0	1.0	0.0	1.0	0.995902	0.995902
10	310.0	0.0	1.0	1.0	0.0	1.0	0.995902	0.995902

Most már behelyettesíthetünk a 3.12, illetve a 3.11 képletbe, mert minden szükséges adat megvan hozzá, így $c_m = 12,081$ és

$$W_{i2}^{1329} = \begin{cases} -1, & \text{ha } i = 1, 2, \dots, 1329 \\ 12,081, & \text{különben.} \end{cases}$$

Ezek után ki tudjuk számolni V_m értékét is $m = 1329$ esetén a 3.13 alapján (ehhez is készítettem egy segédtáblázatot pythonban), az eredmény

$V_{1329} = 4,206$ lett. Ezzel sajnos még nem tudjuk azt mondani, hogy kész a próbastatisztikánk, mert az összes lehetséges m -re vett V_m -ek szuprémuma adja meg azt (a 3.14 alapján): ez $V = 4,662$ lett ($m = 1814$ mellett).

Most már csak a p -értéket kellene meghatározni a bizonyos bootstrap módszer segítségével (a programozás rész ezen a linken keresztül érhető el: <https://github.com/Beszt1/szakdoga/blob/main/bootstrap.ipynb>). Ehhez 30-szor visszatevéssel mintát veszünk, és mindegyik alkalommal kiszámoljuk a V próbastatisztikát (többször is lehetne, úgy még pontosabb lenne az eredmény, de a számítógépem ennyit tudott lefuttatni értelmes idő alatt). Majd ennek segítségével a p -érték a következő lesz: $p_2 = \frac{|V_{bootstrap} \geq V|}{30} = 0,2$, ahol a számlálóban a $V_{bootstrap}$ a visszatevésees minták próbastatisztikáját jelöli.

A fenti eredményeket és a 3.30-at felhasználva: p – érték = $0,0253 + 0,2 \cdot (1 - 0,0253) = 0,220$. Ez az érték nagyobb, mint $\alpha = 0,05$, így a nullhipotézist elfogadjuk, ami azt jelenti, hogy nincs jelentős bizonyíték arra, hogy a hormonkezelés befolyásolná a túlélés valószínűségét.

4. fejezet

Biztonsági hipotézis

Ezt az alkalmazási területet Clarke, Newman (2006) és van Dijk (2007) vetette fel, majd Farrell, Tilley és Tseloni (2010, 2011) fejlesztette tovább. Arra lettek figyelmesek a kutatók, hogy az 1990-es évek elejétől számos iparosodott országban jelentősen csökkent a bűnözés - az Egyesült Államokban az emberölések száma 40%-kal kevesebb lett. De más országokban is megfigyelhető ez az örömteli változás a bűnözés terén, például Angliában és Walesben a járműlopások száma 65%-kal csökkent.

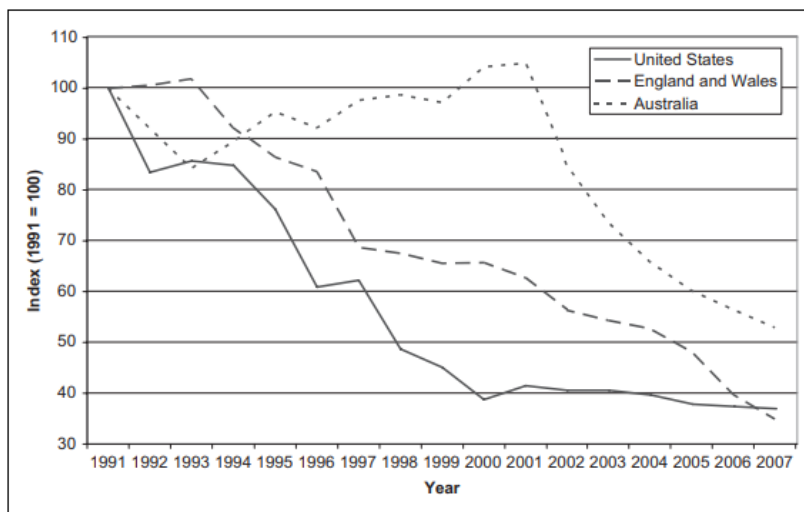
A kutatók azt az állítást vizsgálták, hogy a múlt évtizedekben a bűnözés csökkenésének a fő oka lehetett-e a biztonsági felszerelések mennyiségének és minőségének változása a nyugati országokban. Természetesen nem minden bűncselekménytípus népszerűsége hanyatlott az utóbbi időszakban (például a telefonlopások száma nőtt), de a többségükönél teljesül az észrevétel.

Azért nem egyértelmű alapvetően, hogy mi a fő mozgatórugója a fenti tendenciának, mert számos más dolog is változott az elmúlt években. Például a megnövekedett börtönbüntetés, az illegális kábítószerpiacok is szerepet játszhattak a bűnözések csökkenésében, vagy Amerikában akár az abortusz legalizálása is közrejátszhatott. Mégis a biztonsági eszközök erős pozitív hatását a fent említett kutatók több szemszögből is belátták.

A következő alfejezetekben az autólopásokat és a betöréseket vizsgáljuk meg részletesebben, mert ezek más bűncselekményekre is hatással lehetnek (például a lopott autók segítségével a kábítószerpiacokra könnyebb eljutni). Ráadásul az elmúlt évtizedekben minden nyugati országban drasztikusan megnőtt a vagyon elleni bűncselekmények megelőzésére irányuló intézkedések száma.

4.1. Gépjárműlopás

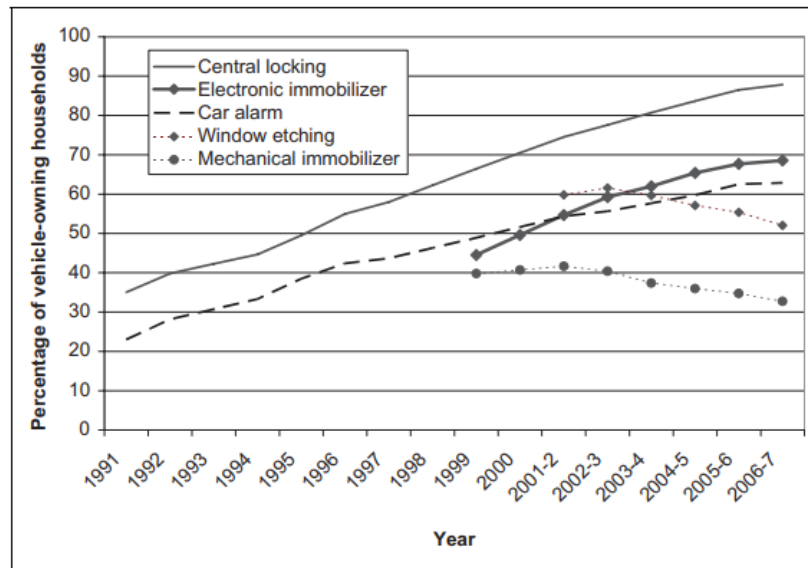
Az Egyesült Államokban 1991 óta eltelt évtizedben 60%-kal csökkent a gépjárműlopások száma. Két évvel később Angliában és Walesben is hasonló visszaesés történt, majd 2001-től Ausztráliában is megfigyelhető ez a tendencia (a korábbi években a gépjárműlopások száma nőtt). Az alábbi ábra mutatja az országokban lévő csökkenéseket (forrás: [11]):



4.1. ábra. Gépjárműlopások száma - USA, Anglia és Wales, Ausztrália

A biztonsági hipotézishez 4 biztonsági eszközt vizsgálunk: mechanikus indításgátlót, elektronikus indításgátlót, riasztót és központi zárat. Ezek a berendezések különböző módon védik az autót. Például az indításgátló nem nehezíti meg az autóból történő lopást, viszont a járműlopásra hatással van. Ezzel szemben a riasztók pedig pont fordított módon hatnak. A következőkben a környezet biztonságát nem vesszük figyelembe, de az is releváns tényező. Ezen felül pedig más biztonsági eszközöket sem vizsgálunk, mert az újabb eszközök nagyobb sikert arattak.

Angliában és Walesben a személygépkocsikra szerelt biztonsági berendezések számának változásait a következő ábra mutatja (forrás: [11]):



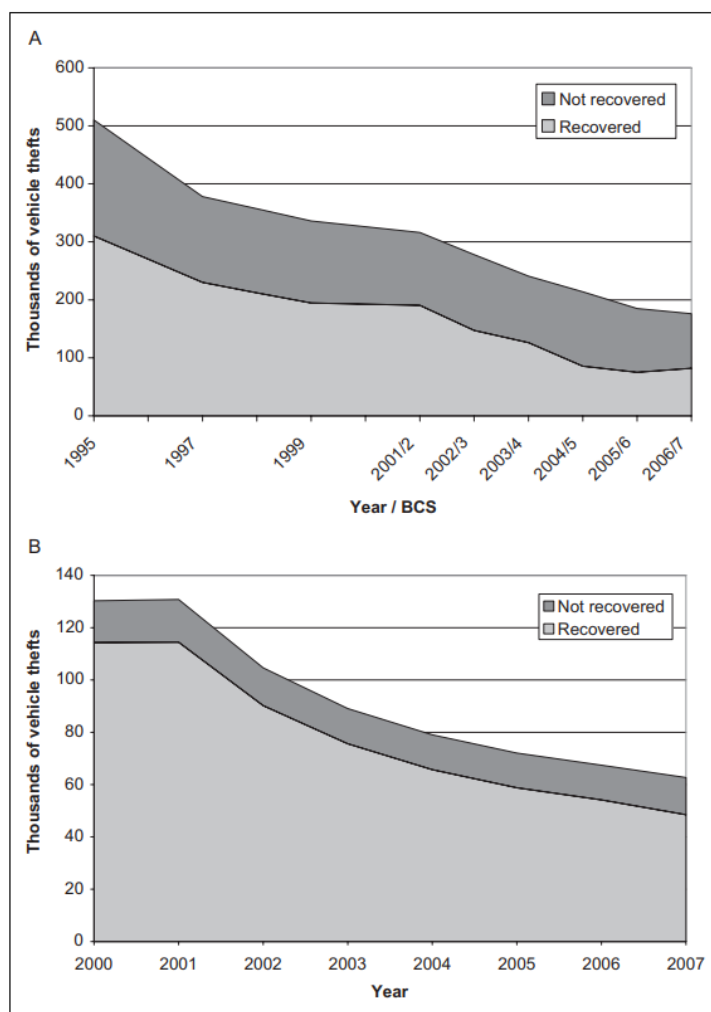
4.2. ábra. Biztonsági rendszerek változása a háztartásokban - Anglia és Wales

Jól látszik, hogy több biztonsági eszköz is egyre elterjedtebbé vált az évek folyamán. Azonban a mechanikus indításgátlóval felszerelt autók száma és a szélvédő gravírozás is csökkent, valószínűleg a népesség körében ezek nem tűntek annyira hatékony eszközöknek. Még megfigyelhető, hogy 2006 és 2007 között több mint kétszer annyi autó rendelkezett elektronikus indításgátlóval, mint mechanikussal.

Az utóbbi két ábrából azt állapíthatjuk meg, hogy Angliában és Walesben ahogy nőtt a járműbiztonság elterjedtsége, úgy csökkent az autólopások száma. Viszont ehhez az állításhoz további bizonyítékok szükségesek.

Ausztráliában az indításgátlókra összpontosítottak. Nyugat-Ausztrália tartományi kormánya már 1997-től támogatta az elektromos indításgátló beszerelését, 1999-ben pedig kötelezővé tette a szabványos indításgátlókat. 2001-ben egész Ausztráliában bevezették, hogy minden új járműbe kötelezően be kell szerelni az ausztrál szabvány szerinti indításgátlót. Ezen felül pedig igyekeztek ösztönözni az embereket, hogy a régebbi autók is rendelkezzenek ezzel a biztonsági eszközzel. Az intézkedések hatásosak voltak, mert 2000-ben a járművek 27,4%-a rendelkezett indításgátlóval, míg 2004-ben már 64,7%-ra nőtt ez az érték. Ráadásul 2001-től a járműlopások száma is elkezdett csökkenni.

Az autólopásoknak van két kategóriája: tartós és ideiglenes. A két típus között az a különbség, hogy az ideiglenes lopásnál a járművet szállításra, száguldozásra szeretné használni a tolvaj, míg tartós lopásnál továbbértékesítés, illetve bizonyos alkatrészek megszerzése a cél. A következő ábra ezeknek a bűncselekményeknek a számának a változását mutatja (forrás: [11]):



4.3. ábra. Autólopások számának változása - Anglia és Wales, Ausztrália

Az ábra alapján (fenti rész Angliára és Walesre vonatkozik, az alsó rész pedig Ausztráliára) a fő csökkenés az ideiglenes járműlopásoknál következett

be. Ennek az lehet a magyarázata, hogy az amatőr autótolvajok kevésbé képesek vagy motiváltak arra, hogy jobb biztonsági eszközökkel felszerelt járműveket törjenek fel csak azért, hogy például száguldhassanak. Észrevehető, hogy a tartós autólopások száma is csökkent, ami arra utal, hogy a hivatásos autótolvajokat is meg lehet állítani a biztonsági intézkedésekkel.

Ha a biztonság javulása tényleg hatással van a gépjárműlopásokra, akkor a lopott autók átlagéletkorának növekednie kellett, mivel az újabb járműveket egyre nehezebb ellopni. Ezt Ausztráliában 2000 és 2007 között leellenőrizték, és valóban kijött a lopott járművek előregedése.

A járművekbe való betörés főleg a jobb zárok következtében csökkent Angliában, Walesben és Ausztráliában is. Azonban a központi zárat ki lehet játszani azzal, ha a tolvajok a kulcsot lopják el. De nincs bizonyíték arra, hogy a kulcsos betörések száma nagyon megnőtt volna. A zárfeltöréseken kívül az ablaktörések száma is látványosan csökkent 1995 óta.

Összességében a különböző biztonsági eszközök eltérő hatást fejtenek ki. A nyomkövető és az elektronikus indításgátló jobban befolyásolja az autólopást, míg a központi zár és a riasztó az autókból történő lopás ellen is hatékonyan véd. Így akkor a legnagyobb a védőhatás mértéke, ha több biztonsági eszközt is beszerelnek a járművekbe.

A fenti hipotézisvizsgálatban 3 ország adatait elemeztük (Anglia, Wales és Ausztrália), ami nagyon kevés ahhoz, hogy általánosítani tudjuk a feltevést. Magyarországon is szerettem volna megnézni a biztonsági hipotézist, de sajnos nem találtam hozzá alkalmas adathalmazt. Ráadásul kérdőív segítségével sem sikerült gyűjteni megfelelő információkat, mert az emberek nem emlékeznek arra, hogy az ellopott autójukban milyen biztonsági rendszerek voltak beszerelve, illetve mennyi ideig volt a birtokukban.

4.2. Lakásbetörés

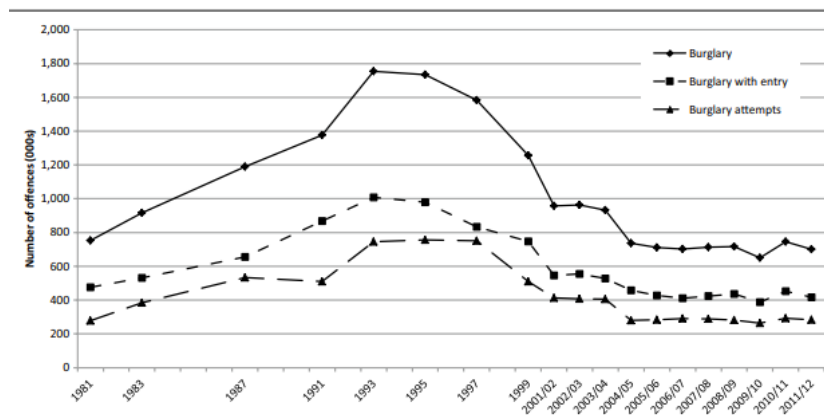
Ha a biztonság fontos szerepet játszott a lakásbetörések számának csökkenésében, akkor a következő változások várhatóak:

- háztartások biztonsági szintje növekszik,
- egyre több háztartás rendelkezik biztonsági eszközzel,
- nő a hatékonyabb biztonsági rendszerek telepítése,

- egyre hatékonyabbak a biztonsági berendezések,
- azok a betöréses lopások csökkennek nagy mértékben, amelyeknél a biztonsági rendszert le kell küzdeni.

4.2.1. Anglia és Wales

A következő ábra Angliában és Walesben történő betörések számát mutatja (forrás: [12]):



4.4. ábra. Betörések száma - Anglia és Wales

Megfigyelhető, hogy 1993 és 2011-2012 között 64%-kal csökkent a regisztrált betörések száma. A legnagyobb visszaesés 1997 és 2001-2002 között volt, átlagosan évi 10%-kal csökkent az esetek száma. 2004-2005 óta viszonylag stabil a tendencia, évente körülbelül 700000 betörés történt. A betörők „találathi aránya” (a behatolással elkövetett betörések aránya a betörések teljes számához viszonyítva) 1981 óta 59% körül ingadozott, és a legmagasabb értéket (63%) a növekvő bűnözés időszakában (1981 és 1991) érte el. Ez az arány 2004-2005 óta eléggé stabilan 60%-on maradt.

A következő megállapításokhoz egy bűnügyi felmérést használunk (CSEW - Crime Survey for England and Wales), amely Anglia és Wales felnőtt lakosságától gyűjtött adatokat a bűncselekményekkel kapcsolatos tapasztalatokról. A betörés áldozatait megkérdezték az események részleteiről, a betörés idején beszerelt biztonsági eszközökről. Ezenkívül egy véletlenszerűen kiválasztott almintát kérdeztek meg a lakásukba felszerelt biztonsági eszközökről.

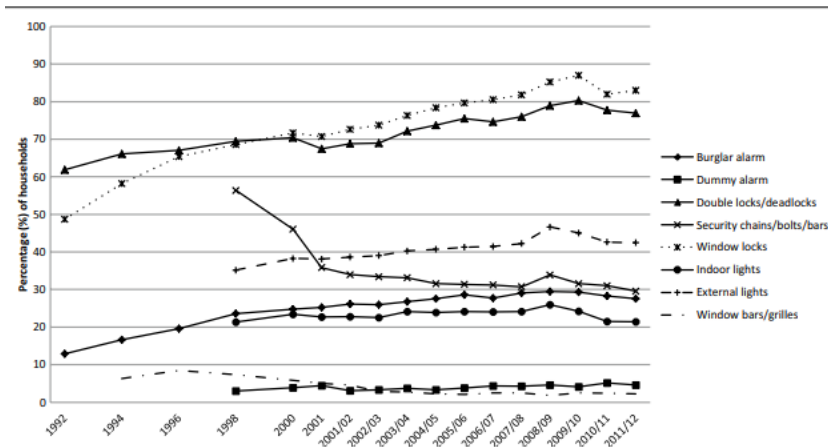
Eme adatok segítségével a betörési kockázat és a biztonsági rendszerek közötti kapcsolatot meg lehet vizsgálni.

Először a biztonsági berendezésekkel nem rendelkező háztartásokat hasonlítjuk össze az adott biztonsági eszközzel vagy eszközkombinációval rendelkező háztartások kockázatával. Az eredményül kapott mérőszám (SPF - security protection factor) azt mutatja meg, hogy mennyivel nehezebben sebezhető meg egy olyan célpont, ami rendelkezik adott biztonsági rendszerrel.

Ezután az alfejezet elején megfogalmazott 5 hipotézist teszteljük.

- Növekedett-e a háztartások biztonsági szintje?

1992 és 2011-2012 között általánosan nőtt a betöréses lopást megelőző biztonsági berendezésekkel felszerelt háztartások aránya, amit az alábbi ábra támaszt alá (forrás: [12]):



4.5. ábra. Biztonsági rendszerek elterjedése a háztartásokban

Lineáris regresszió segítségével meg lehet határozni, hogy milyen típusú változás (növekedés, csökkenés vagy stagnálás) figyelhető meg a különböző biztonsági berendezések elterjedésében. Ennek a módszernek az a lényege, hogy megkeresi az ismert pontjainkra (x , y koordinátát tartalmaznak) a lehető legjobban illeszkedő lineáris egyenest. Az alábbi állítás segítségével meghatározhatóak a lineáris egyenes paraméterei:

4.2.1. Állítás. Legyenek $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ adott számpárok.

Azokat az a és b együtthatókat keressük, melyre a

$$h^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (ax_i + b))^2$$

minimális. Ennek megoldása:

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{k=1}^n (x_k - \bar{x})^2}$$

és

$$\hat{b} = \bar{y} - \hat{a}\bar{x}.$$

Hipotézisvizsgálatra van szükségünk ahhoz, hogy megállapítsuk, hogy valójában történt-e szignifikáns változás a berendezések elterjedésében az idő előrehaladtával - ehhez egyoldali t-próbát alkalmazunk. A következő táblázat a fenti diagramról leolvasható adatokból készült python segítségével (a kód az alábbi linken keresztül megtalálható: https://github.com/Beszti1/szakdoga/blob/main/linearis_regresszio.ipynb):

Berendezés	Regressziós egyenlet	Hipotézis	t értéke	Döntés
Duplázár	$y = 0.8341x - 1598.8341$	$H_0 : a \leq 0, H_1 : a > 0$	10.536	Elutasítjuk
Biztonsági lánc	$y = -1.3296x + 2701.3893$	$H_0 : a \geq 0, H_1 : a < 0$	-3.885	Elutasítjuk
Ablakzár	$y = 1.5949x - 3120.5949$	$H_0 : a \leq 0, H_1 : a > 0$	11.659	Elutasítjuk
Riasztó	$y = 0.7367x - 1450.2367$	$H_0 : a \leq 0, H_1 : a > 0$	8.571	Elutasítjuk
Álriasztó	$y = 0.1728x - 341.7858$	$H_0 : a \leq 0, H_1 : a > 0$	3.912	Elutasítjuk
Beltéri lámpa	$y = 0.1124x + 202.3195$	$H_0 : a \leq 0, H_1 : a > 0$	0.999	Elfogadjuk
Kültéri lámpa	$y = 0.6260x - 1214.3266$	$H_0 : a \leq 0, H_1 : a > 0$	7.012	Elutasítjuk
Ablakrácsok	$y = -0.3049x + 615.1810$	$H_0 : a \geq 0, H_1 : a < 0$	-7.028	Elutasítjuk

4.1. táblázat. Lineáris regressziós eredmények

Minden sorban a nullhipotézis az alapján lett választva, hogy a lineáris regressziós egyenletben a meredekség pozitív vagy negatív előjelű. Ez azért praktikus, mert így könnyen el tudjuk majd dönteni, hogy volt-e

tényleges változás az adott berendezés népszerűségében. Egyoldali t-próbát végzünk, aminek a képlete a következő, ha $y = \hat{a}x + \hat{b}$ alakú egyenes:

$$t = \hat{a} \frac{\sqrt{(n-2) \sum_{i=1}^n (X_i - \bar{X})^2}}{\sqrt{\sum_{i=1}^n (Y_i - \hat{a}X_i - \hat{b})^2}}. \quad (4.1)$$

A konkrét értékeket behelyettesítve kapjuk meg a t-próba értékét, amelyek a táblázat utolsó előtti oszlopában is láthatóak. A kritikus érték ($n-2$ szabadsági fok és $\alpha = 0,05$ szignifikanciaszint mellett) abszolútértéke minden berendezés esetében 1,8 (csak a második tizedesjegytől kezdve térnek el a számok). A hipotézisvizsgálat eldöntéséhez össze kell hasonlítani a kapott t -t és t_{krit} -et. Ha H_0 -t $a \leq 0$ -nak választottuk, akkor ezt jobboldali t-próbának hívjuk, és abban az esetben utasítjuk el a H_0 -t, ha $t > t_{krit} = 1,8$. Viszont, ha H_0 $a \geq 0$ -val egyezik meg, akkor baloldali t-próbánk van, és $t < t_{krit} = -1,8$ esetében tudjuk elutasítani a nullhipotézist.

Az eredmények alapján a vizsgált időszakban az ablakzár beszerelésénél történt a legnagyobb növekedés (körülbelül 50%-ról 80% fölé emelkedett az arány). A beltéri lámpáknál történt a legkevesebb változás, mert ott fogadtuk el egyedül a nullhipotézist - nincs jelentős bizonyítékunk arra, hogy egyre népszerűbbé vált volna ez a biztonsági berendezés. Jelentős csökkenés csak a biztonsági láncnál figyelhető meg, aminek az lehet az oka, hogy nem nyújtott eléggé hatáson védelmet, bizonyos szerszámok segítségével könnyen szétszedhető a lánc. A fenti 4.5 diagram annyiban hiányos, hogy a kombinációkat és az eszközök számát nem mutatja. Ezért a leggyakrabban használt kombinációk össze lettek gyűjtve. Azt vették észre, hogy csak azon kombinációk aránya csökkent a háztartásokra nézve, amelyek tartalmazták a biztonsági láncot, ami nem meglepő. 2011-2012-ig a legnépszerűbb kombináció az ajtó- és ablakzár tartalmazta. Tehát az adatok alapján kijelenthető, hogy növekedett a vizsgált időszakban a háztartások biztonsági szintje.

- Csökkent-e a biztonsági rendszerrel nem rendelkező háztartások aránya? Nőtt-e a relatív betörési kockázatuk?

Az adatok alapján jelentősen csökkent a biztonsági eszközökkel nem

rendelkező háztartások aránya (1992 és 1996, valamint 2008-2009 és 2011-2012 között több mint 70%-os volt a visszaesés). Nem teljesen pontos a megállapítás, mert a felmérés listáján szereplő biztonsági eszközök száma növekedett az 1990-es években, így nagyobb esély lett arra, hogy valamelyik berendezés megtalálható a válaszadó háztartásában. Viszont az biztos, hogy egyre több háztartásban megtalálható valamilyen biztonsági eszköz.

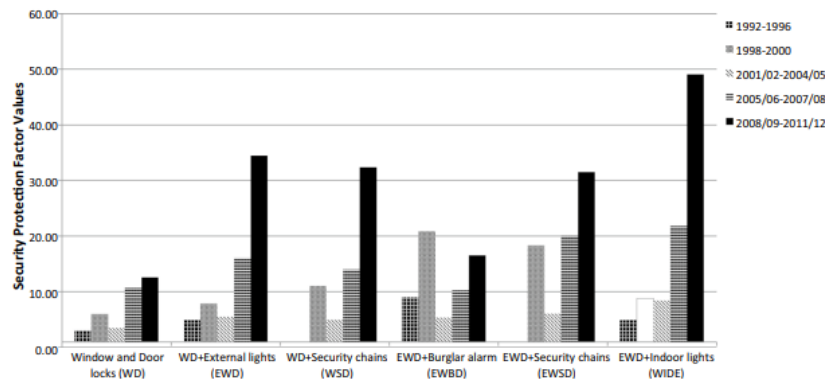
Megállapították azt is, hogy míg az 1990-es évek elején csak kétszer akkora valószínűséggel törtek be olyan háztartásokba, amelyek nem voltak felszerelve biztonsági eszközökkel, addig 2010 környékére 8-szoros esélye lett ennek. Ebből levonható az a következtetés, hogy az előbb említett háztartások relatív betörési kockázata 4-szeresére nőtt a bűnözés csökkenése alatt.

- Növekedett-e a hatékonyabb biztonsági berendezések telepítése?

Vannak olyan biztonsági eszközök, amelyek az SPF mutató alapján hatékonyabban csökkentik a betörés kockázatát, de például a riasztó nem ilyen (valószínűleg az az oka ennek, hogy a betörők a riasztó esetében azt remélik, hogy értékesebb tárgyat szerezhetnek meg). Általánosságban megfogalmazható, hogy minél több biztonsági eszköz van felszerelve, annál védettebb lesz a háztartás (4-nél több eszköznél már annyira nem észrevehető az állítás). A leghatékonyabb kombinációkkal (ablak- és ajtózárok, ablak- és ajtózárok mozgásérzékelő lámpákkal, illetve még az utóbbiakhoz hozzáadva az időzítővel ellátott belső világítást) rendelkező háztartások aránya majdnem megduplázódott. Továbbá az eszközkombinációk széleskörű elterjedésével szemben a biztonsági rendszerrel nem rendelkező háztartások mellett az egyetlen eszközzel rendelkező háztartások száma is csökkent.

- Javult-e a biztonsági berendezések hatékonysága?

A CSEW felmérés a biztonsági berendezések minőségéről és arról, hogy az eszközök használatban voltak-e a betörésnél, nem gyűjtött adatokat. Viszont a hatékonyság növekedése is egy kulcskérdés a biztonsági hipotézis szempontjából. Emiatt a különböző eszközkombinációk SPF-jeit kiszámolták az egyes időintervallumokra (1992-1996, 1998-2000, 2001-2005, 2005-2008 és 2008-2012-re). Az alábbi ábra mutatja az eredményeket (forrás: [12]):



4.6. ábra. Eszközkombinációk SPF-je

Boldogan megállapíthatjuk, hogy egy kivétellel (EWD + riasztó = mozgásérzékelős lámpa, ablak- és ajtózárok és a riasztó együttes kompozíciója) mindegyik kombináció SPF mutatója erősen növekedett. Ebből arra következtethetünk, hogy mind a hatékonyság, mind a minőség javult. A riasztó 2000-es évektől kezdődően nem annyira hatékony, aminek az okát már korábban vizsgáltuk, ezért csökken a fent említett kilógó kombinációnak az SPF mutatója.

Tehát összességében a biztonsági eszközök egyre jobban elterjedtek, és a riasztó kivételével jobban működnek a betörések elhárításában.

- Nagyobb mértékben csökkentek-e azok a betörések, ahol voltak biztonsági berendezések?

Megkülönböztetünk erőszakos (zár vagy ablak feltörése...) és erőszak nélküli betöréseket (nyitva hagyott ablak vagy ajtó, kulccsal való behatolás...). Azt figyelték meg, hogy az erőszakos betörések száma jelentősen csökkent (1990-es évek és a 2000-es évek között több mint 70%-os visszaesést tapasztaltak). Ezzel szemben az erőszak nélküli behatolással elkövetett betörések aránya nem nagyon változott.

A biztonsági eszközök általában akkor lépnek működésbe, ha senki sincsen otthon. Emiatt az üresen maradt lakásoknál nagyobb mértékű csökkenést várunk a betörések terén. Ezt a sejtést a CSEW adatai alátámasztják.

A fenti pontok alapján a biztonsági hipotézis valóban teljesül. Azonban

azt nem szabad elfelejteni, hogy csak egy területen, egy felmérés alapján vizsgáltuk meg a hipotézist.

4.2.2. Egyesült Amerikai Államok

Most egy másik földrészen is megvizsgáljuk, hogy mi lehet a betörések csökkenésének a fő oka ([14] és [15] források alapján készült ez az alfejezet). Az Egyesült Amerikai Államokban 1980 és 2019 között több mint 80%-kal visszaesett a betörések száma.

Bostoni tanulmány megállapította, hogy az 1980-as évek betörőinek jelentős többsége fiatal (25 év alatti) és nem képzett. Ebből arra lehet következtetni, hogy elég sok háztartásból hiányzott a jó minőségű biztonsági rendszer. Másrészt pedig többen gondatlanul nyitva felejtették az ajtót, ablakot vagy a garázst, ami miatt szintén sokkal könnyebb volt elvinni értéktárgyakat idegen tulajdonú házból. Ezeket a következtetéseket alátámasztják az 1980-as évekbeli adatok, miszerint a betörések 42%-a erőszak nélkül történt. Továbbá a háztartások mindössze 7%-a rendelkezett riasztóval, amelyek az esetek több mint a felében nem működtek megfelelően.

New York államban végzett 1983-84-es felmérés a másik oldalt is alátámasztotta, miszerint a jó minőségű biztonsági rendszer hatékony védelmet biztosít. Nem sok háztartásban volt fellelhető az utóbbi rendszer, de ahol igen, ott sikeresen megakadályozta a betörési kísérleteket.

A biztonsági rendszerek hatékony alkalmazására bizonyíték, hogy az évek során az idősebb betörők aránya nőtt, a fiataloké csökkent (a kevésbé tapasztaltakat jobban elriasztják a jó minőségű biztonsági eszközök). Tehát arra a következtetésre juthatunk, hogy manapság a sikeres betörések sokkal több időt és erőfeszítést vesznek igénybe. Ezt alátámasztja az is, hogy egyre elterjedtebbek a biztonsági eszközök, és a hatékonyságuk is folyamatosan nőtt.

Továbbra is igaz, hogy még mindig nem elég sok helyen vizsgáltuk meg a lakásbetörésekhez kapcsolódó biztonsági hipotézist, hogy általánosítani tudjunk. Viszont a fent említett területeken megfigyelt hipotézisvizsgálatok elég jó kiindulást adnak.

Irodalomjegyzék

- [1] Backhausz Ágnes jegyzetei - <https://backhauszagi.web.elte.hu/gyak/>
- [2] Hans-Peter Beck-Bernholdt, Hans-Hermann Dubben, A tojást rakó kutyá, Magyar Könyvklub (2001)
- [3] <https://www.kaggle.com/datasets/utkarshx27/breast-cancer-dataset-used-royston-and-altman?resource=download> - letöltés időpontja 2024.10.15.
- [4] Klein, J. P., Moeschberger, M. L., Klein, J. P., & Moeschberger, M. L. (2003). Hypothesis testing. Survival analysis: techniques for censored and truncated data, 201-242.
- [5] <https://hu.wikipedia.org/wiki/T%C3%BA1%C3%A91%C3%A9s-anal%C3%ADzis>
- [6] Dormuth, I., Liu, T., Xu, J., Yu, M., Pauly, M., & Ditzhaus, M. (2022). Which test for crossing survival curves? A user's guideline. BMC medical research methodology, 22(1), 34.
- [7] Li, H., Han, D., Hou, Y., Chen, H., & Chen, Z. (2015). Statistical inference methods for two crossing survival curves: a comparison of methods. PLoS One, 10(1), e0116774.
- [8] Qiu, P., & Sheng, J. (2008). A two-stage procedure for comparing hazard rate functions. Journal of the Royal Statistical Society Series B: Statistical Methodology, 70(1), 191-208.
- [9] Fleming, T. R., & Harrington, D. P. (2013). Counting processes and survival analysis. John Wiley & Sons.

- [10] Vaart, A. W. (1998). Asymptotic Statistics.
- [11] Farrell, G., Tseloni, A., Mailley, J., & Tilley, N. (2011). The crime drop and the security hypothesis. *Journal of Research in Crime and Delinquency*, 48(2), 147-175.
- [12] Tseloni, A., Farrell, G., Thompson, R., Evans, E., & Tilley, N. (2017). Domestic burglary drop and the security hypothesis. *Crime Science*, 6, 1-16.
- [13] Soto, H. (2021). Testing the security hypothesis as explanation for Chilean burglary trends (Doctoral dissertation, UCL (Univeristy College London)).
- [14] Farrell, G. Why has burglary declined by 80 percent across four decades in the United States? Evidence relating to the security hypothesis. https://ideas.repec.org/p/osf/socarx/c78wz_v1.html - letöltés időpontja: 2025.01.22.
- [15] Farrell, G. (2022). Forty years of declining burglary in the United States: Explanation and evidence relating to the security hypothesis. *Security Journal*, 35(2), 444-462.

Alulírott **Bajnok Eszter** nyilatkozom, hogy szakdolgozatom elkészítése során az alább felsorolt feladatok elvégzésére a megadott MI alapú eszközöket alkalmaztam:

Feladat	Felhasznált eszköz	Felhasználás helye	Megjegyzés
Python kód megírása a példákhoz	GPT-4o	példáknál	Csak hibüzenetek esetén, illetve, ha valahol elakadtam a kód írásában, akkor kértem segítséget az AI-tól.
Latexben írás	GPT-4o	képleteknél	Amikor nem jutott eszembe, hogy hogyan kell kalapot csinálni a betű fölé, hogyan kell táblázatot készíteni, akkor használtam AI-t.

A felsoroltakon túl más MI alapú eszközt nem használtam.

