Sportrekordok matematikai modellezése

Szakdolgozat

Kovács Dávid

Matematika BSC
Alkalmazott matematikus szakirány

Témavezető:

Dr. Zempléni András
tanszékvezető, egyetemi docens
Valószínűségelméleti és Statisztika Tanszék
Eötvös Loránd Tudományegyetem, Természettudományi Kar

Eötvös Loránd Tudományegyetem
Természettudományi Kar
2019

Köszönetnyilvánítás

Köszönöm témavezetőmnek, Dr. Zempléni Andrásnak a témajavaslatot és a szakdolgozat megírásában nyújtott folyamatos segítségét. Hálás vagyok, amiért számtalan alkalommal átnézte a munkámat és rengeteg hasznos tanáccsal látott el.
Köszönöm továbbá édesanyámnak, hogy az évek alatt mindig mellettem állt és támogatott mindenben. Nélküle nem juthattam volna el idáig.

Tartalomjegyzék

Bevezetés **4**

1. Extrémérték-elméleti alapok **6**
1.1. A maximum eloszlásbeli konvergenciája 6
1.2. Maximum vonzási tartomány 8
1.2.1. Fréchet eloszlás (Φ_α) maximum vonzási tartománya . . . 9
1.2.2. Weibull eloszlás (Ψ_α) maximum vonzási tartománya . . . 10
1.2.3. A Gumbel eloszlás (Λ) vonzási tartománya 10
1.3. Küszöbmeghaladás 10

2. A rendezett minta aszimptotikus viselkedése **13**
2.1. Határvalószínűségek 13
2.2. Határeloszlások 17

3. A rekordok gyakorisága **21**
3.1. A rekordok száma véges kezdőszeletben 21
3.2. A rekordok aszimptotikus gyakorisága 24
3.3. Várakozási idő rekordok között 26

4. Sportrekordok modellezése **28**
4.1. Síkfutás 28
4.1.1. A tendencia javulása 28
4.1.2. A paraméterek becslése 30
4.1.3. Usain Bolt eredményei 35
4.1.4. A modell ellenőrzése 36
4.1.5. Rekordok a jövőben 40
4.1.6. A 400 méteres síkfutás 43
4.2. Távolugrás 43
4.2.1. A helyparaméter alakulása 44
4.2.2. Rekordok a jövőben 46

Összefoglalás **47**

Bevezetés

Ennek a szakdolgozatnak a célja különféle atlétikai számok rekorderedményeinek modellezése. A 100 méteres síkfutás különös figyelmet kap a negyedik fejezetben sorra kerülő modellezés során. Megpróbáljuk előrejelezni, hogy a Usain Bolt által beállított $9,58$ másodperces rekord mikor dől meg. Valamint becslést adunk arra, hogy mi az emberiség által elérhető legjobb eredmény, vagyis mennyi az a minimális idő, mely alatt lefutható a 100 méteres táv. Ezen felül valamivel kisebb részletességgel a 400 méteres síkfutásról és a távolugrásról is szó esik az említett negyedik fejezetben.

Ehhez azonban először meg kell teremteni a kellő matematikai hátteret. Ez három fejezeten át történik. Az első fejezetben az extrémérték-elmélethez kapcsolódó alapfogalmak kerülnek ismertetésre. Az itt leírtak alapkövét képezik a későbbiekben elmondottaknak.

Az extrémérték-elmélet a maximumok eloszlásával foglalkozik, azonban gyakran pontosabb becslést kaphatunk a rendezett minta több elemének felhasználásával. Ennek fényében a második fejezet a rendezett minták aszimptotikus eloszlását tárgyalja.

A harmadik fejezetben pedig azt vizsgáljuk, hogy független, azonos eloszlású, abszolút folytonos valószínűségi változókból milyen gyakran adódnak rekordok. Ez is egy fontos téma, hiszen a bizonyításra kerülő állítások segítségével bizonyos esetekben könnyen elvethetjük a független, azonos eloszlású minta hipotézisét. Más esetekben pedig épphogy jó közelítésel szolgálhat ez a modell.

Ezek után tehát már minden szükséges ismeret rendelkezésünkre áll majd, és rátérhetünk a modellezésre a negyedik fejezetben. Itt a célunk az lesz, hogy az évenkénti legjobb eredmények tendenciájára paraméteres becslést adjunk, és ezáltal előrejelzést próbálunk majd adni az eredmények jövőbeli alakulásáról. A fejezet egyrészt a [2] könyvben és [7] cikkben leírtakra, másrészt egyéni munkára épül. A [2] könyv és [7] cikk tartalmazza a modellezés alapgondolatait. Az egyéni munkát pedig főként ezen alapgondolatok és az első három fejezetben leírtak összehangolása alkotja. Vagyis például a két szóban forgó forrásban látott eljárások kiegészítésre, továbbfejlesztésre kerülnek az eddig leírtak segítségével. A

továbbfejlesztést egyrészt a felhasznált évek számának növelése adja. Ugyanis a [7] cikkel ellentétben nem csak a rekordot beállító éveket tekintjük, hanem minden rendelkezésünkre álló évből használunk fel adatot. A [2] könyv pedig ugyan szintén minden rendelkezésre álló évet felhasznál, mégis kiadása óta 18 év eltelt, ez több mint másfélszeres növekedés az elemezhető évek számában. Ezenfelül a felhasznált adatok jellegében is változást eszközölünk, hiszen a két megjelölt forrással ellentétben rendelkezésünkre áll számos évnek a legjobb eredményekből kapott rendezett mintája. Így lehetőségünk nyílik a második, a rendezett minták aszimptotikus eloszlásáról szóló fejezetben leírtak felhasználására. Ezen kívül szintén az egyéni munka részét képezi annak a felvetésnek a vizsgálata, miszerint Usain Bolt eredményei teljes mértékben külön kezelendőek, mivel túl jó időket produkált ahhoz, hogy elfogadható módon be lehessen illeszteni a majd megalkotott modellbe. A 400 méteres síkfutás és a távolugrás vizsgálata is saját munkán alapszik, ugyanis ezeket az atlétikai számokat nem modellezte a két említett forrás.

1. fejezet

Extrémérték-elméleti alapok

Ebben a fejezetben a továbbiakban szükséges, az extrémérték-elmélet témaköréhez tartozó alapismeretek kerülnek ismertetésre. Gyakran feldolgozásra kerülő témakörről lévén szó, ebben a szakdolgozatban a szóban forgó alapok precíz felépítése nem elsődleges szempont. Ennek megfelelően ebben a fejezetben a tételek bizonyítás nélkül kerülnek közlésre. A fejezet mélységében és felépítésében főként az [1] és [2] könyvekhez illeszkedik. Az itt le nem írt bizonyítások pedig megtalálhatóak például a [4] és [5] könyvekben.

1.1. A maximum eloszlásbeli konvergenciája

Ebben az alfejezetben a mintaelemek maximumának aszimptotikus viselkedésével fogunk foglalkozni. Ez kézenfekvő vizsgálódási terület a sportrekordok elemzése szempontjából, hiszen egy adott sport esetén az egy évben adódó legjobb eredmény sok eredmény maximuma (vagy éppen minimuma, de (-1)-gyel való szorzással a minimum visszavezethető a maximumra, úgyhogy a továbbiakban a maximum vizsgálatára szorítkozunk). Ideális esetben feltételezhetjük, hogy kellően nagy az évenkénti minta, és így kellően jó becsléssel szolgálnak az aszimptotikus tulajdonságokra támaszkodó módszerek.

1.1.1. Definíció. *Legyenek $X_1, X_2, \ldots$ valószínűségi változók, és legyen $n \in \mathbb{N}^+$. Ekkor $M_n := max(X_1, \ldots, X_n)$.*

1.1.2. Definíció. *Legyen X valószínűségi változó F eloszlásfüggvénnyel, ekkor F jobb végpontján az $x_F := sup\{x \in \mathbb{R} : F(x) < 1\} \in (-\infty, +\infty]$ értéket értjük.*

1.1.3. Állítás. *Legyenek $X_1, X_2, \ldots$ független valószínűségi változók F elosz-*

lásfüggvénnyel. Ekkor

$$F_{M_n} = F^n \longrightarrow \begin{cases} \mathbb{1}_{[x_F,+\infty)} & ,ha\ x_F < \infty \\ 0 & ,ha\ x_F = \infty \end{cases}.$$

Ez a gyakorlatban nem túl hasznos konvergencia. Így tehát az az ötlet adódik, hogy valamilyen $(a_n)_{n\in\mathbb{N}^+} \subset \mathbb{R}^+$ és $(b_n)_{n\in\mathbb{N}^+} \subset \mathbb{R}$ választással $F_{M_n}(x)$ helyett $F_{(M_n-b_n)/a_n}(x) = F^n(a_nx + b_n)$-t vizsgáljuk, azaz M_n-hez megfelelő normálókonstansokat keresünk, annak reményében, hogy így nem elfajuló határeloszlás adódik.

1.1.4. Tétel. *Legyenek Y,$Y_1, Y_2 \ldots$ valószínűségi változók , valamint $(a_n)_{n\in\mathbb{N}^+} \subset \mathbb{R}^+$ és $(b_n)_{n\in\mathbb{N}^+} \subset \mathbb{R}$ normáló konstansok, melyekre*

$$(Y_n - b_n)/a_n \xrightarrow{d} Y.$$

Ekkor ha $(a'_n)_{n\in\mathbb{N}^+} \subset \mathbb{R}^+$ és $(b'_n)_{n\in\mathbb{N}^+} \subset \mathbb{R}$ olyan normáló konstansok, melyekre

$$\lim_{n\to\infty} a_n/a'_n = a \in \mathbb{R}_0^+ \text{ és } \lim_{n\to\infty} (b_n - b'_n)/a'_n = b \in \mathbb{R},$$

akkor

$$(A_n - b'_n)/a'_n \xrightarrow{d} aY + b.$$

Az 1.1.4. tételnek a mondanivalója az, hogy ha normáló konstansok segítségével eloszlásbeli konvergenciát kapunk, akkor a kapott határeloszlás lineáris transzformáltjait is megkaphatjuk határértékként a megfelelő normáló konstansokkal. A következő tételt joggal nevezik az extrémérték-elmélet alaptételének.

1.1.5. Tétel (Fisher–Tippett). *Legyenek $X_1, X_2, \ldots$ független valószínűségi változók F eloszlásfüggvénnyel. Tegyük fel, hogy $\exists\,(a_n)_{n\in\mathbb{N}^+} \subset \mathbb{R}^+$ és $(b_n)_{n\in\mathbb{N}^+} \subset \mathbb{R}$, melyre $(M_n - b_n)/a_n \xrightarrow{d} Y$, ahol Y nem elfajuló, abszolút folytonos valószínűségi változó G eloszlásfüggvénnyel. Ekkor $\exists\,\alpha \in \mathbb{R}^+$, hogy G a következő három eloszlás valamelyikének lineáris transzformáltja:*

1. $\Phi_\alpha(x) := \exp(-x^{-\alpha})\mathbb{1}_{[0,\infty)}(x)$
2. $\Psi_\alpha(x) := \exp\big(-(-x)^\alpha\big)\mathbb{1}_{[-\infty,0)}(x) + \mathbb{1}_{[0,\infty)}(x)$
3. $\Lambda(x) := \exp\big(-\exp(-x)\big)$.

A három típust rendre Fréchet, Weibull, és Gumbel eloszlásnak nevezik. Az 1.1.5. tételben megjelenő 3 eloszlásfüggvényt, és azoknak lineáris transzformáltjait extrémérték-eloszlásnak nevezik. A tételt a normáló konstansok ismerete nélkül is jól fel tudjuk használni céljainkra. Tekintsünk egy adott sportot, és

egy adott évet. Abban az évben kellően sok, n atléta eredményéből született az évi maximum. A pontos értékét persze nem tudjuk n-nek. Mi M_n eloszlását szeretnénk valahogy becsülni. Feltételezzük, hogy n kellően nagy, így az aszimptotikus becslés már elég pontos.

Tehát

$$\mathbb{P}(M_n < x) = \mathbb{P}(\frac{M_n - b_n}{a_n} < \frac{x - b_n}{a_n}) \approx G(\frac{x - b_n}{a_n}), \tag{1.1}$$

ahol G extrémérték-eloszlás, a_n, b_n a megfelelő normáló konstansok. Vegyük észre, hogy $H(x) = G(\frac{x-b_n}{a_n})$ is egy extrémérték-eloszlás. Tehát M_n eloszlását egy extrémérték-eloszlással tudjuk közelíteni, melynek paramétereit nem ismerjük, de persze megbecsülhetjük. Így az eddigiekkel egyeztetve azt kapjuk, hogy egy adott év (vagy egyéb kellően hosszú időintervallum) legjobb eredményei közelítőleg valamilyen extrémérték-eloszlás szerint oszlanak el, ezen extrémérték-eloszlás paramétereinek becslése központi szerepet fog játszani a 4. fejezet során. Viszont érdemes észben tartani, hogy ez a gondolatmenet akkor helyes, ha létezik eloszlásbeli határérték a megfelelő normáló konstansokkal. Erről a kérdésről, vagyis az 1.1.5 tétel alkalmazhatóságáról röviden az 1.2. alfejezetben esik szó.

A gyakorlatban sokszor kényelmetlennek bizonyul az, hogy egy jelenség modellezésénél a három lehetséges határeloszlás közül választanunk kell egyet, vagy ha nem tudunk választani, akkor mindegyiket bele kell foglalnunk külön az eljárásunkba. Ennek köszönhetően felmerült az igény egy olyan modell megalkotására, mely magában foglalja mindhárom határeloszlást. A széles körben elterjedt standard reprezentáció az általánosított extrémérték-eloszlás (GEV).

1.1.6. Definíció. *Azt mondjuk, hogy az Y valószínűségi változó standard GEV eloszlású $\xi \in \mathbb{R}$ alakparaméterrel, ha eloszlásfüggvénye*

$$G_\xi(x) = \exp\big(-(1+\xi x)^{-1/\xi}\big)\mathbb{1}_{A(\xi)}(x) + \mathbb{1}_{(-\infty,0)\times(-1/\xi,\infty)}(\xi, x),$$

ahol $A(\xi) := \{x \in \mathbb{R} : 1+\xi x \geq 0\}$, és a $\xi = 0$ eset úgy értendő, hogy a megfelelő $\xi \to 0$ határértéket vesszük.

A standard GEV eloszlást transzformálva $\mu \in \mathbb{R}$ helyparaméterrel és $\sigma \in \mathbb{R}^+$ skálaparaméterrel kapjuk meg a (μ, σ, ξ) paraméterű GEV eloszlást, melyet $GEV(\mu, \sigma, \xi)$ jelöl. Ez a fajta felírás tehát az összes esetet tartalmazza egy formulában, ami gyakorlati szempontból igencsak kézenfekvő.

1.2. Maximum vonzási tartomány

Ebben az alfejezetben röviden ismertetjük a maximum vonzási tartomány fogalmát, melynek segítségével valamivel nagyobb rálátásunk adódhat arra, hogy alkalmazható-e 1.1.5. a sportrekordok modellezésében.

1.2.1. Definíció. *Azt mondjuk, hogy az L nemnegatív függvény lassú változású, ha* $\lim_{x\to\infty} \frac{L(ax)}{L(x)} = 1$.

1.2.2. Definíció. *Azt mondjuk, hogy az f nemnegatív függvény reguláris változású, ha $\forall a > 0$ esetén $g(a) := \lim_{x\to\infty} \frac{f(ax)}{f(x)}$ véges és nem nulla.*

1.2.3. Tétel (Karamata). *Legyen f reguláris változású függvény.*
Ekkor $f(x) = x^\alpha L(x)$, valamilyen $\alpha \in \mathbb{R}$ számra, ahol L lassú változású.

Ekkor α-t a reguláris változás indexének nevezzük, és azt, hogy az f reguláris változású függvény α indexszel a következőképpen jelöljük: $f \in \mathscr{R}_\alpha$.

1.2.4. Definíció. *Azt mondjuk, hogy F eloszlásfüggvény a H eloszlásfüggvény maximum vonzási tartományában van $\big($jelölés: $F \in MDA(H)\big)$, ha $\exists\, (a_n)_{n\in\mathbb{N}^+} \subset \mathbb{R}^+$ és $(b_n)_{n\in\mathbb{N}^+} \subset \mathbb{R}$ normáló konstansok, melyekre*

$$F^n(a_n x + b_n) \overset{n\to\infty}{\longrightarrow} H(x).$$

1.2.1. Fréchet eloszlás (Φ_α) maximum vonzási tartománya

Ahogy azt az 1.1.4. tételben láttuk a vonzási tartomány szempontjából csak az alakparaméternek, azaz α-nak van jelentősége. Ezt érdemes észben tartani a következő tétel értelmezésénél.

1.2.5. Tétel. *Legyen F eloszlásfüggvény. Ekkor*

$$F \in MDA(\Phi_\alpha) \iff F \in \mathscr{R}_{-\alpha}.$$

Ez magával vonja, hogy $F \in MDA(\Phi_\alpha)$ esetben $x_F = \infty$. Így ha mondjuk olyan sportrekordokat modellezünk, amelyekben egy adott távot kell minél rövidebb idő alatt teljesíteni, és ezt úgy tesszük, hogy (-1)-gyel szorozzuk az eredményeket, és a maximumeloszlást modellezzük, akkor Fréchet-eloszlás nem adódhat, hiszen egészen biztosan $x_F < 0$ ebben az esetben. De ha például távolugrást modellezünk, akkor sem ésszerű feltételezés az $x_F = \infty$, hiszen különböző fizikai és biológiai okokból kiindulva felső korlát szabható az eredményeknek.

Tehát, amikor $GEV(\mu, \sigma, \xi)$ eloszlással modellezünk sportrekordokat, akkor $\xi \leq 0$-nak kell adódnia, ellenkező esetben problémás lehet a modellünk. Ez egy jó kiindulási alap lesz az eljárásunk ellenőrzésében. Ennek ellenére megjegyzendő, hogy ugyan elméleti szempontból nem indokolt a pozitív alakparaméter megjelenése, mégis előfordulhat, hogy valamilyen pozitív alakparaméterű GEV eloszlás modellezi jól az adott sport eredményeit. Ugyanis előfordulhat, hogy egy ilyen GEV eloszlású valószínűségi változó kellően nagy valószínűséggel csak valamilyen észszerűnek tűnő korlát alatti értékeket vesz fel.

1.2.2. Weibull eloszlás (Ψ_α) maximum vonzási tartománya

Természetesen itt is csak az alakparaméternek van jelentősége a vonzási tartomány vizsgálatánál.

1.2.6. Tétel. *Legyen F eloszlásfüggvény. Ekkor*

$$F \in MDA(\Psi_\alpha) \iff x_F < \infty, G \in \mathscr{R}_{-\alpha},$$

ahol $G(x) = \overline{F}(x_F - 1/x)$.

Az $x_F < \infty$ tulajdonság alapján a Weibull eloszlás egy jó lehetséges jelölt sportrekordok modellezésére. Persze 1.1.5. alkalmazhatóságára még mindig nincs garanciánk, de az érzésünk az lehet, hogy, ha minden atléta mondjuk évi legjobb teljesítménye külön-külön F eloszlásfüggvényű, akkor az 1.2.6.-beli feltételek teljesülhetnek F-re. Az erről való megbizonyosodásra ugyan nyilván nincs precíz mód, de bizonyos értelemben pozitív visszajelzés, hogy a feltételek teljesíthetőnek tűnnek. Így arra a következetetésre jutunk, hogy érdemes lehet valóban az aszimptotikus modellel próbálkozni a sportrekordok modellezése során.

1.2.3. A Gumbel eloszlás (Λ) vonzási tartománya

Mivel a 4. fejezetben numerikus módszerekkel fogunk maximum likelihood becslést számolni a megfelelő GEV paraméterekre, nyilván pontosan $\xi = 0$ érték nem fog előállni. De persze a nagyon kicsi abszolútértékű ξ értékek arra utalhatnak, hogy a kapott extrémérték-eloszlás igazából Gumbel, csak a különféle pontatlanságokból adódóan nem jöhet ki kereken $\xi = 0$.

A most elmondottak miatt és amiatt, hogy a Gumbel eloszlás vonzási tartományát nehezebb jellemezni, csak annyit jegyzünk meg itt, hogy $x_F < \infty$ és $x_F = \infty$ értékek is adódhatnak $F \in MDA(\Lambda)$ esetén. Ebből következően, ha esetleg olyan kicsi ξ értéket kapunk, hogy azt sejtjük, hogy valójában Gumbel eloszlással van dolgunk, nem feltétlenül kell megijednünk, hiszen az x_F végességének szempontjából még nem adódott hiba a modellben. A modell helyességéről majd úgyis különféle módszerekkel próbálunk meggyőződni, így nem is lenne most érdemes a vonzási tartományok elméletét hosszabban ismertetni.

1.3. Küszöbmeghaladás

1.3.1. Definíció. *Legyen X valószínűségi változó F eloszlásfüggvénnyel és x_F jobb végponttal. Ekkor $u \in (-\infty, x_F)$ esetén*

$$F_u(x) := \mathbb{P}(X - u < x | X \geq u)\mathbb{1}_{[0,\infty)}(x).$$

Ekkor azt mondjuk, hogy F_u az u-beli küszöbmeghaladás eloszlásfüggvénye.

1.3.2. Definíció. *Legyen X valószínűségi változó x_F jobb végponttal és legyen $u \in (-\infty, x_F)$ adott. Ekkor*

$$e(u) := \mathbb{E}(X - u | X \geq u).$$

Ekkor azt mondjuk, hogy $e(u)$ az u-beli küszöbmeghaladás várható értéke.

1.3.3. Definíció. *Azt mondjuk, hogy Z valószínűségi változó standard általánosított Pareto eloszlású ξ alakparaméterrel, ha eloszlásfüggvénye*

$$H(x) = \big(1 - (1 + \xi x)^{-1/\xi}\big) \mathbb{1}_{[0,\infty) \cap A(\xi)}(x),$$

ahol a $\xi = 0$ eset a $\xi \to 0$ határérték vételével értendő.

A standard általánosított Pareto-eloszlást transzformálva $\nu \in \mathbb{R}$ helyparaméterrel és $\beta \in \mathbb{R}^+$ skálaparaméterrel kapjuk meg a (ν, β, ξ) paraméterű általánosított Pareto-eloszlást, melyet $GPD(\nu, \beta, \xi)$ jelöl. Az általánosított Pareto-eloszlás bevezetését a következő tétel indokolja (ismét figyelembe véve az 1.1.4. tétel következményeit).

1.3.4. Tétel.

$F \in MDA(G_\xi) \iff \exists b : (-\infty, x_F) \to \mathbb{R}^+ \textit{mérhető} : \forall x \in A(\xi) :$
$\lim\limits_{u \to x_F - 0} F_u\big(b(u)x\big) = 1 - (1 + \xi x)^{-1/\xi},$

ahol G_ξ standard GEV eloszlásfüggvény ξ alakparaméterrel, és a $\xi = 0$ eset most is a megfelelő határérték vételével értendő.

Ez a tétel a gyakorlatban a következőképpen használható. Tegyük fel, hogy u kellően nagy ahhoz, hogy 1.3.4. -beli aszimptotikus tulajdonságok már jó közelítéssel szolgáljanak. Ekkor

$$F_u(x) = F_u\left(b(u)\frac{x}{b(u)}\right) \approx \left(1 + \xi\frac{x}{b(u)}\right)^{-1/\xi} = H(x),$$

ahol H egy $GPD\big(0, b(u), \xi\big)$ eloszlásfüggvény. Azaz a küszöbtúllépés $GPD\big(0, b(u), \xi\big)$ eloszlással közelíthető. Persze tipikusan a szóban forgó b függvényt sem ismerjük, szóval rögzített u esetén $b(u)$-ra is egy becsülendő paraméterként tekintünk.

A küszöbtúllépés vizsgálata a sportrekordok elemzése esetében is kézenfekvő tud lenni. Előfordul ugyanis, hogy csak egy bizonyos küszöböt túllépő eredményeket jegyeznek fel, így a küszöbtúllépési modell alkalmazására szorulunk.

Emellett a küszöbtúllépési modell lehetővé teszi, hogy több adattal tudjunk dolgozni, mint csak az éves maximumot vizsgálva. Például a 4. fejezetben bemutatásra kerülő, az alakparaméterre vonatkozó becslésnek is a GPD modell áll a hátterében.

2. fejezet

A rendezett minta aszimptotikus viselkedése

Ebben a fejezetben a rendezett mintával kapcsolatos határértéktételek és egyéb állítások kerülnek tárgyalásra. A fejezet az [1] és [3] könyvekben leírtakra támaszkodik elsősorban. Lévén, hogy a későbbiekben bemutatott modellezési módszerek alapjául a rendezett minta aszimptotikus tulajdonságainak kihasználása fog szolgálni, a bizonyításokra nagyobb hangsúly helyeződik ebben a fejezetben.

2.1. Határvalószínűségek

Először – az alapvető definíciók ismertetése után – vizsgáljuk a határértékét néhány a rendezett mintával kapcsolatos valószínűségnek. Az így adódó tételek megfelelő átalakítása fogja eredményezni a határeloszlás-tételeket.

2.1.1. Definíció. *Legyen $k \in \mathbb{N}^+$ adott, valamint legyen H olyan halmaz, melyre $k \leq |H| < \infty$. Ekkor $\max_k H$ jelölje a H halmaz k. legnagyobb elemét (ha több is van, akkor mindegy melyiket választjuk).*

2.1.2. Definíció. *Legyenek $X_1, X_2, \ldots$ valószínűségi változók, valamint legyen $k, n \in \mathbb{N}^+, k \leq n$ adott. Ekkor $X_{k,n} := \max_k\{X_1, \ldots, X_n\}$.*

2.1.3. Definíció. *Legyenek $X_1, X_2, \ldots$ valószínűségi változók, továbbá legyen adott $(u_n)_{n\in\mathbb{N}^+}$ sorozat. Ekkor*

$$B_n(u) := \sum_{j=1}^{n} \mathbb{1}_{\{X_j \geq u_n\}}.$$

Ekkor tehát $B_n(u)$ az u_n szintet meghaladó mintaelemek száma az első n megfigyelés között.

2.1.4. Tétel (Poisson határvalószínűség). *Legyenek $X_1, X_2, \dots$ független valószínűségi változók F eloszlásfüggvénnyel, továbbá legyen $(u_n)_{n\in\mathbb{N}^+}$ olyan sorozat, melyre teljesül az $n\overline{F}(u_n) \longrightarrow \tau$ feltétel, valamilyen $\tau \in [0,\infty]$ számra. Ekkor $\forall k \in \mathbb{N}_0$ esetén*

$$\mathbb{P}\big(B_n(u) \le k\big) \longrightarrow e^{-\tau} \sum_{j=0}^{k} \frac{\tau^j}{j!}, \tag{2.1}$$

ahol a jobb oldalt $\tau = 0$ esetén 1-nek, míg $\tau = \infty$ esetben 0-nak értjük. Megfordítva, ha (2.1) teljesül valamilyen $k \in \mathbb{N}_0$ esetén, akkor $n\overline{F}(u_n) \to \tau$ (és így (2.1) teljesül $\forall k \in \mathbb{N}_0$ számra).

Bizonyítás.

Tegyük fel, hogy $n\overline{F}(u_n) \longrightarrow \tau$.

A $\tau \in (0,\infty)$ eset egyszerűen a Poisson nevéhez fűződő ismert tételen múlik, azt használjuk ki, hogy $B_n(u) \xrightarrow{d} Y \sim Poi(\tau)$. Ez pedig éppen (2.1).

Ha $\tau = 0$:

$$1 \ge \mathbb{P}\big(B_n(u) \le k\big) \ge \mathbb{P}\big(B_n(u) = 0\big) = \big(1 - \overline{F}(u_n)\big)^n =$$
$$= \left(1 - \frac{1}{n}\big(n\overline{F}(u_n)\big)\right)^n \longrightarrow e^0 = 1.$$

Így rendőrelvet alkalmazva valóban $\mathbb{P}\big(B_n(u) \le k\big) = 1$.

Ha $\tau = \infty$:

Mivel $n\overline{F}(u_n) \longrightarrow \infty \ \ \forall \theta > 0$ esetén $n\overline{F}(u_n) \ge \theta$ elég nagy $n \ge n_0$ esetén megfelelő $n_0 \in \mathbb{N}^+$ küszöbbel.

Így $n \ge n_0$ esetén:

$$\mathbb{P}\big(B_n(u) \le k\big) = \sum_{j=0}^{k} \left[\binom{n}{j} \big(\overline{F}(u_n)\big)^j \big(1 - \overline{F}(u_n)\big)^{n-j}\right] \le$$
$$\le \sum_{j=0}^{k} \left[\binom{n}{j} \left(\frac{\theta}{n}\right)^j \left(1 - \frac{\theta}{n}\right)^{n-j}\right] =$$
$$= \sum_{j=0}^{k} \left[\frac{\theta^j}{j!} \left(1 - \frac{\theta}{n}\right)^n \left(1 - \frac{\theta}{n}\right)^{-j} \frac{n(n-1)\dots(n-j+1)}{n^j}\right] \overset{n\to\infty}{\longrightarrow} e^{-\theta} \sum_{j=0}^{k} \frac{\theta^j}{j!}.$$

Itt kihasználtuk, hogy a binomiális eloszlás eloszlásfüggvénye pontonként monoton csökken a valószínűséget megadó paraméter növekedésével. Így

$$0 \leq \limsup_{n\to\infty} \mathbb{P}\big(B_n(u) \leq k\big) \leq e^{-\theta} \sum_{j=0}^{k} \frac{\theta^j}{j!} \overset{\theta\to\infty}{\longrightarrow} 0.$$

Vagyis valóban $\lim_{n\to\infty} \mathbb{P}\big(B_n(u) \leq k\big) = 0$.

A fordított irány bizonyításához indirekt tegyük fel, hogy (2.1) teljesül valamilyen $l \in \mathbb{N}_0$ számra, de $n\overline{F}(u_n) \nrightarrow \tau$. Ekkor $\exists\, (u_{n_k}) \subset (u_n)$, melyre $n\overline{F}(u_{n_k}) \longrightarrow \tau'$, valamilyen $\tau \neq \tau' \in [0,\infty]$ számra (ha korlátos $n\overline{F}(u_n)$, akkor van konvergens részsorozat, ha nem korlátos van végtelenbe tartó részsorozat). Ekkor a tétel már bebizonyított másik iránya szerint $B_{n_k}(u) \xrightarrow{d} Y \sim Poi(\tau')$, ami pedig ellentmondás.

□

Az eddigiekből könnyen megkapható a következő tétel.

2.1.5. Tétel (A k. legnagyobb elem határvalószínűsége). *Legyenek $X_1, X_2, \ldots$ független valószínűségi változók F eloszlásfüggvénnyel, továbbá legyen $(u_n)_{n\in\mathbb{N}^+}$ olyan sorozat, melyre teljesül az $n\overline{F}(u_n) \longrightarrow \tau$ feltétel, valamilyen $\tau \in [0,\infty]$ számra. Ekkor*

$$\lim_{n\to\infty} \mathbb{P}(X_{k,n} < u_n) = e^{-\tau} \sum_{j=0}^{k} \frac{\tau^j}{j!},$$

ahol a jobb oldal értelmezése a (2.1) egyenletnél említett módon történik.

Bizonyítás. A szóban forgó határvalószínűség meghatározásában 2.1.4. jól felhasználható. Ugyanis

$$\lim_{n\to\infty} \mathbb{P}(X_{k,n} < u_n) = \lim_{n\to\infty} \mathbb{P}\big(B_n(u) \leq k-1\big) = e^{-\tau} \sum_{j=0}^{k-1} \frac{\tau^j}{j!}.$$

□

2.1.6. Tétel (A rendezett minta együttes határvalószínűsége). *Legyen $k \in \mathbb{N}^+$ és $X_1, X_2, \ldots$ független valószínűségi változók F eloszlásfüggvénnyel adottak, továbbá $(u_n^{(k)})_{n\in\mathbb{N}^+} \leq \ldots \leq (u_n^{(1)})_{n\in\mathbb{N}^+}$ legyenek olyan sorozatok, melyekre $\forall j = 1, \ldots, k$ esetén*

$$\lim_{n\to\infty} n\overline{F}(u_n^{(j)}) = \tau_j \in [0,\infty]. \tag{2.2}$$

Ekkor $\forall l_1, \ldots, l_k \in \mathbb{N}$ esetén

$$\lim_{n\to\infty} \mathbb{P}\big(B_n(u^{(1)}) = l_1, B_n(u^{(2)}) = l_1 + l_2, \ldots, B_n(u^{(k)}) = l_1 + \ldots + l_k\big) =$$
$$= \frac{\tau_1^{l_1}}{l_1!} \frac{(\tau_2 - \tau_1)^{l_2}}{l_2!} \cdots \frac{(\tau_k - \tau_{k-1})^{l_k}}{l_k!} e^{-\tau_k},$$

ahol a jobb oldal $\tau_k = \infty$ esetben 0-ként értendő.

Bizonyítás.

Legyen $p_{n,j} = \overline{F}(u_n^{(j)})$. Így ha $n \in \mathbb{N}^+$ rögzített

$$\mathbb{P}\big(B_n(u^{(1)}) = l_1, B_n(u^{(2)}) = l_1 + l_2, \ldots, B_n(u^{(k)}) = l_1 + \ldots + l_k\big) =$$
$$= \binom{n}{l_1} p_{n,1}^{l_1} \binom{n-l_1}{l_2} (p_{n,2} - p_{n,1})^{l_2} \cdots$$
$$\cdots \binom{n-l_1-\ldots-l_{k-1}}{l_k} (p_{n,k} - p_{n,k-1})^{l_k} (1 - p_{n,k})^{n-l_1-\ldots-l_k}.$$

Itt azt használtuk ki, hogy a szóban forgó valószínűség leszámolható oly módon, hogy kiválasztjuk $X_1, \ldots, X_n$ közül az $u_n^{(1)}$-et túllépőket, ezek persze $u_n^{(2)}, \ldots, u_n^{(k)}$ mindegyikét túllépik. Majd kiválasztjuk azokat, melyek $u_n^{(1)}$-et nem lépik túl, de $u_n^{(2)}$-t igen, melyek persze $u_n^{(3)}, \ldots, u_n^{(k)}$ mindegyikét túllépik . Így folytatva végül azokat választjuk ki, melyek $u_n^{(1)}, \ldots, u_n^{(k)}$ közül egyiket sem lépik túl.

Ha $\tau_k < \infty$ ($\tau_1 \leq \ldots \leq \tau_k$ miatt ez azzal ekvivalens, hogy $\tau_1, \ldots, \tau_k < \infty$), akkor (2.2) alapján

$$\binom{n}{l_1} p_{n,1}^{l_1} \sim \frac{(np_{n,1})^{l_1}}{l_1!} \longrightarrow \frac{\tau_1^{l_1}}{l_1!}$$

$$\ldots$$

$$\binom{n-l_1-\ldots-l_{k-1}}{l_k} (p_{n,k} - p_{n,k-1})^{l_k} \sim \frac{(np_{n,k} - np_{n,k-1})^{l_k}}{l_k!} \longrightarrow \frac{(\tau_k - \tau_{k-1})^{l_k}}{l_k!}$$

$$(1 - p_{n,k})^{n-l_1-\ldots-l_k} \sim \left(1 - \frac{np_{n,k}}{n}\right)^n \longrightarrow e^{-\tau_k}.$$

Ezzel megkaptuk a bizonyítani kívánt formulát.

Ha $\tau_k = \infty$, akkor 2.1.4. alkalmazásával

$$\mathbb{P}\big(B_n(u^{(1)}) = l_1, B_n(u^{(2)}) = l_1 + l_2, \ldots, B_n(u^{(k)}) = l_1 + \ldots + l_k\big) \leq$$
$$\leq \mathbb{P}\big(B_n(u^{(k)}) = l_1 + \ldots + l_k\big) \longrightarrow 0.$$

□

2.2. Határeloszlások

Most, hogy a kellő határvalószínűséggel kapcsolatos tételek kidolgozásra kerültek, áttérhetünk a nagyobb gyakorlati jelentőséggel bíró határeloszlás-tételekre.

2.2.1. Definíció. *Legyen H extrémérték-eloszlás, $k \in \mathbb{N}^+$. Ekkor*

$$H^{(k)}(x) := H(x) \sum_{j=0}^{k-1} \frac{\big(-\log H(x)\big)^j}{j!}, \qquad x \in \mathbb{R}$$

ahol most a $log(0) = 0$ konvencióval élünk.

Ekkor 1.1.5. és 2.1.5. közvetlen következményeként a következő tétel adódik.

2.2.2. Tétel (A k. legnagyobb elem határeloszlása). *Legyen H extrémérték-eloszlás, $F \in MDA(H)$ $(a_n)_{n\in\mathbb{N}^+} \subset \mathbb{R}^+$ és $(b_n)_{n\in\mathbb{N}^+} \subset \mathbb{R}$ normáló konstansokkal. Továbbá legyenek $X_1, X_2, \ldots$ független valószínűségi változók F eloszlásfüggvénnyel. Ekkor $\forall k \in \mathbb{N}^+$ esetén*

$$\lim_{n\to\infty} \mathbb{P}\Big((X_{k,n} - b_n)/a_n < x\Big) = H^{(k)}(x). \tag{2.3}$$

Megfordítva, ha valamilyen $k \in \mathbb{N}^+$ és G esetén

$$\lim_{n\to\infty} \mathbb{P}\Big((X_{k,n} - b_n)/a_n < x\Big) = G(x),$$

akkor $G(x) = H^{(k)}(x)$, valamilyen H extrémérték-eloszlással, valamint ekkor $\forall k \in \mathbb{N}^+$ esetén teljesül (2.3).

Az eddigieket felhasználva a következő tételt kapjuk a rendezett minta együttes határeloszlására.

2.2.3. Tétel (Az együttes határeloszlás létezése). *Legyen H extrémérték-eloszlás, $F \in MDA(H)$, $(a_n)_{n\in\mathbb{N}^+} \subset \mathbb{R}^+$ és $(b_n)_{n\in\mathbb{N}^+} \subset \mathbb{R}$ normáló konstansokkal, $X_1, X_2, \ldots$ független valószínűségi változók F eloszlásfüggvénnyel. Ekkor $\exists Y$ nem elfajuló k dimenziós valószínűségi vektorváltozó, melyre*
$\Big((X_{j,n} - b_n)/a_n\Big)_{j=1}^{k} \xrightarrow{d} Y.$

Bizonyítás.

Legyen $(x_1, \ldots, x_k) \in \mathbb{R}^k$ tetszőleges. Továbbá

$$u_n^{(1)} := a_n x_1 + b_n, \ldots, u_n^{(k)} := a_n x_k + b_n.$$

Ekkor $F \in MDA(H)$ miatt $\forall x \in \mathbb{R}$ esetén

$$(a_n x + b_n)\overline{F}(a_n x + b_n) \longrightarrow \tau(x) \in [0, \infty].$$

Vagyis

$$u_n^{(1)}\overline{F}(u_n^{(1)}) \longrightarrow \tau(x_1), \ldots, u_n^{(k)}\overline{F}(u_n^{(k)}) \longrightarrow \tau(x_k).$$

Ekkor

$$\mathbb{P}(X_{1,n} < u_n^{(1)}, \ldots, X_{k,n} < u_n^{(k)}) =$$
$$= \mathbb{P}\big(B_n(u^{(1)}) = 0, B_n(u^{(2)}) \leq 1, \ldots, B_n(u^{(k)}) \leq k-1\big).$$

Ezt a valószínűséget fel tudjuk írni olyan valószínűségek összegeként, mint amilyenek a 2.1.6. tételben jelennek meg. És mindegyik tagra külön alkalmazva a 2.1.6. tételbeli formulát megkapjuk a megfelelő határeloszlást. □

A bizonyítás során tehát megadtuk a szóban forgó Y valószínűségi vektorváltozó eloszlásfüggvényét. Kisebb k értékek esetén valóban átlátható formula adódik így az eloszlásfüggvényre. A továbbiakban viszont lesz példa olyan esetre, ahol $k = 10$ választással kerülnek modellezésre sportrekordok, ez már 10! tagot eredményezne, ami nyilván nem ésszerű. Továbbá a maximum likelihood becslésben a sűrűségfüggvény fog kelleni, így érdemesebb inkább a sűrűségfüggvény szempontjából megközelíteni Y-t. Szerencsére a sűrűségfüggvényre jobban használható formulát kapunk.

2.2.4. Állítás. *Legyen $X_1, X_2, \ldots, X_n$ független valószínűségi változó F eloszlásfüggvénnyel, valamint legyen $k \in \mathbb{N}^+, k \leq n$ adott. Ekkor $\left(X_{j,n}\right)_{j=1}^{k}$ együttes sűrűségfüggvénye*

$$f_{k,n}(x) = F^{n-k}(x_k) \prod_{j=1}^{k} \Big[(n-j+1)f(x_j)\Big] \mathbb{1}_{D(k)}(x),$$

ahol $D(k) = \Big\{(x_1, \ldots, x_k) \in \mathbb{R}^k : x_1 < \ldots < x_k\Big\}$

2.2.5. Tétel. *Legyenek $X_1, X_2, \ldots$ független valószínűségi változók F eloszlásfüggvénnyel, és f sűrűségfüggvénnyel. Továbbá legyen Y olyan H eloszlásfüggvényű, h sűrűségfüggvényű valószínűségi változó, melyre*

$\exists\,(a_n)_{n\in\mathbb{N}^+}\subset\mathbb{R}^+, (b_n)_{n\in\mathbb{N}^+}\subset\mathbb{R}$, *hogy*

$$\lim_{n\to\infty}\left(\sup\left\{|Q_n(B)-Q_Y(B)| : B\in\mathscr{B}\right\}\right)=0, \tag{2.4}$$

teljesül, ahol Q_n $(M_n-b_n)/a_n$ *eloszlása,* Q_Y Y *eloszlása,* $\mathscr{B}$ *pedig* $\mathbb{R}$ *Borel-halmazait jelöli. Ekkor* Y *extrémérték-eloszlás, továbbá* $\exists\, Z$ k *dimenziós valószínűségi vektorváltozó, melyre* $\left((X_{j,n}-b_n)/a_n\right)_{j=1}^{k}\xrightarrow{d} Z$, *és ezen* Z *sűrűségfüggvénye*

$$f_Z(x)=H(x_k)\prod_{j=1}^{k}\left(\frac{h(x_j)}{H(x_j)}\right)\mathbb{1}_{D(k)}(x).$$

Bizonyítás.

A (2.4) feltétel következménye, hogy $F\in MDA(H)$, hiszen $\left\{(-\infty,c) : c\in\mathbb{R}\right\}\subset\mathscr{B}$, így $(M_n-b_n)/a_n\xrightarrow{d}Y$. Ekkor az 1.1.5. tétel következményeként Y extrémérték-eloszlás. Így a 2.2.3. tétel miatt $\left((X_{j,n}-b_n)/a_n\right)_{j=1}^{k}$ eloszlásban konvergens.

Legyen h_n $(M_n-b_n)/a_n$ sűrűségfüggvénye. Ekkor a (2.4)-beli feltételt alkalmazva $\{h_n>h\}$ és $\{h_n\le h\}$ halmazokra

$$\int_{\mathbb{R}}|h_n-h|=\int_{\{h_n>h\}}|h_n-h|+\int_{\{h_n\le h\}}|h_n-h|=\int_{\{h_n>h\}}(h_n-h)-\int_{\{h_n\le h\}}(h_n-h)\longrightarrow 0.$$

Tehát $\int_{\mathbb{R}}|h_n-h|$ nullsorozat, így van összegezhető részsorozata. Vagyis $\exists\,(m_n)_{n\in\mathbb{N}^+}\subset\mathbb{R}$, melyre

$$\infty>\sum_{n=1}^{\infty}\int_{\mathbb{R}}|h_{m_n}-h|=\int_{\mathbb{R}}\sum_{n=1}^{\infty}|h_{m_n}-h|.$$

Itt kihasználtuk, hogy az integrandus nemnegatív, így a Fubini-tétel miatt megcserélhető a szummázás (számláló mérték szerinti integrálás) és a Lebesgue-mérték szerinti integrálás.

Vagyis $\sum_{n=1}^{\infty}|h_{m_n}-h|$ integrálja véges, így $\sum_{n=1}^{\infty}|h_{m_n}-h|<\infty$ majdnem mindenütt, ebből pedig $h_{m_n}\xrightarrow{\text{mm.}}h$ következik. Ekkor kihasználva, hogy $h_m(x)=ma_mF^{m-1}(a_mx+b_m)f(a_mx+b_m)$, valamint, hogy $F^{m-1}(a_mx+b_m)\longrightarrow H(x)$,

$$ma_mf(a_mx+b_m)\xrightarrow{\text{mm.}}h(x)/H(x)$$

adódik. Ezt a tényt és a 2.2.4. állítást kihasználva $\left((X_{j,m_n}-b_n)/a_{m_n}\right)_{j=1}^{k}$ vektorváltozó h_{k,m_n} sűrűségfüggvényére

$$h_{k,m_n}(x) = a_{m_n} f_{m_n,k}(a_{m_n}x + b_{m_n}) =$$
$$= F^{m_n-k}(a_{m_n}x_k + b_{m_n}) \prod_{j=1}^{k} \Big[(n-j+1)a_{m_n} f(a_{m_n}x_j + b_{m_n})\Big] \mathbb{1}_{D(k)}(x) H(x_k)$$
$$\longrightarrow H(x_k) \prod_{j=1}^{k} \frac{h(x_j)}{H(x_j)} \mathbb{1}_{D(k)}(x) \quad .$$

adódik. Vagyis $\Big((X_{j,m_n} - b_{m_n})/a_{m_n}\Big)_{j=1}^{k}$ sűrűségfüggvénye majdnem mindenütt a tételben szereplő Z valószínűségi vektorváltozó sűrűségfüggvényéhez konvergál. Ekkor a Scheffé-lemma egy könnyen adódó következményeként

$$\Big((X_{j,m_n} - b_{m_n})/a_{m_n}\Big)_{j=1}^{k} \xrightarrow{d} Z. \tag{2.5}$$

Azt pedig láttuk, hogy $\Big((X_{j,n} - b_n)/a_n\Big)_{j=1}^{k}$ eloszlásban konvergens, így (2.5) miatt az eloszlásbeli határértérték csak Z lehet.

□

3. fejezet

A rekordok gyakorisága

Ebben a fejezetben azt vizsgáljuk, hogy milyen gyakran fordulnak elő rekordok, ha egymástól függetlenül, mindig ugyanabból az eloszlásból értékeket származtatunk. Fő célunk, hogy megmutassuk, hogy ilyen körülmények között jelentősen kevesebb rekord adódik, mint azt a sportrekordoknál megszokhattuk, ezzel igazolva, hogy a sportrekordok nem ilyen módon származnak. A fejezetben szereplő állítások, tételek nagy része bizonyítás nélkül megtalálható a [6] könyvben.

3.1. A rekordok száma véges kezdőszeletben

Az egyik legelemibb kérdés az, hogy egy adott kezdőszeletben várhatóan hány rekord adódik. Ezen kérdés körüljárása előtt érdemes megjegyezni, hogy ha független, abszolút folytonos valószínűségi változóink vannak egy közös F eloszlásfüggvénnyel, akkor a szóban forgó kérdés megválaszolása szempontjából nem is számít, hogy konkrétan mi is F. Ennek az az oka, hogy az adott feltételek mellett minden sorrend egyformán valószínű, így a sorrendekkel megfogalmazható tulajdonságok nem függenek F-től. Ez a gondolatmenet a későbbiekben precízzé lesz téve.

3.1.1. Definíció. *Legyen adott $X_1, X_2, \ldots$ valószínűségi változó sorozat. Ekkor k=1 esetén $I_k :\equiv 1$, $2 \leq k \in \mathbb{N}$ esetén $I_k := \mathbb{1}_{\{X_k = X_{1,k}\}}$.*

Tehát I_k egy indikátor valószínűségi változó, mely azt mutatja, hogy X_k rekord-e.

3.1.2. Definíció. *Legyen adott $X_1, X_2, \ldots$ valószínűségi változó sorozat és $n \in \mathbb{N}^+$. Ekkor $R_n := \sum_{k=1}^{n} I_k$.*

Tehát R_n a rekordok számát adja meg az n hosszú kezdőszeletben.

3.1.3. Lemma. *Legyen adott $X_1, X_2, \ldots$ független, azonos eloszlású, abszolút folytonos valószínűségi változó sorozat. Legyen $k, l \in \mathbb{N}^+$, $k < l$, jelölje Π_k az $\{1, 2, \ldots, k\}$ halmaz permutációinak halmazát. Továbbá*

$$\mathscr{F}_k := \big\{\{X_{i_1} < X_{i_2} < \ldots < X_{i_k}\} : (i_1, \ldots, i_k) \in \Pi_k\big\}.$$

Ekkor $H_l^{(r)} := \{X_l = X_{r,l}\}$ és $\mathscr{F}_k$ függetlenek.

Bizonyítás.

Legyen $F \in \mathscr{F}_k$. Vagyis legyen adott egy rögzített sorrendje az első k elemnek. Ekkor $\mathbb{P}(F \cap H_l^{(r)})$ értéke elemi kombinatorikus gondolatmenettel megkapható. Megint azt használjuk ki, hogy minden sorrend egyformán valószínű, egyezések pedig csak 0 valószínűséggel vannak.

Számoljuk le, hány eset van, amikor az első k elem a szóban forgó rögzített sorrendben van, és az l-edik elem r-rekord. Gondoljuk úgy a leszámolásra, hogy mindegyik elemnek adunk egy sorszámot. Az l-edik elem csak az r sorszámot kaphatja. Az első k elemnél csak az számít melyik k darab sorszám van ott, a sorrendjük egyértelmű, azaz a maradék $l-1$ sorszámból választunk ide k darabot, ez $\binom{l-1}{k}$ féleképpen tehető meg. A maradék $l-k-1$ helyre tetszőleges sorrendben mehetnek a sorszámok, ez $(l-k-1)!$ lehetőség. Tehát összesen $\binom{l-1}{k}(l-k-1)! = \frac{(l-1)!}{k!}$ féleképpen áll elő a kívánt sorrend. Vagyis

$$\mathbb{P}(F \cap H_l^{(r)}) = \frac{(l-1)!}{k! \cdot l!} = \frac{1}{k!}\frac{1}{l} = \mathbb{P}(F)\mathbb{P}(H_l^{(r)}).$$

Tehát valóban független $H_l^{(r)}$ és $\mathscr{F}_k$.

□

A 3.1.3. állításból rögtön következik, hogy $H_l^{(r)}$ és $\sigma(\mathscr{F}_k)$ is független. Ennek az a mondanivalója az, hogy egy rekordjellegű esemény az l-edik helyen mindentől független, amit az első k darab esemény sorszámával tudunk megfogalmazni. Ez egy jóval általánosabb eredmény, mint ami nekünk közvetlenül szükséges. De ebből adódóan láthatjuk, hogy ha például R_n mintájára definiáljuk a k-rekordok (az addigi kezdőszeletben k. legnagyobb elemek) $R_n^{(k)}$ számát, vagy a $(-k)$-rekordok (az addigi kezdőszeletben k. legkisebb elemek) $R_n^{(-k)}$ számát, akkor azok is R_n-hez hasonló módon tárgyalhatóak. De egyéb rekordszerű fogalmakat is definiálhatunk, azonban a továbbiakban főként R_n vizsgálatára szorítkozunk. Az eddigiek segítségével nem nehéz a következőt belátni.

3.1.4. Tétel. *Legyen adott $X_1, X_2, \ldots$ független, azonos eloszlású, abszolút folytonos valószínűségi változó sorozat. Ekkor $I_1, I_2, \ldots$ függetlenek. Továbbá $\forall m \in \mathbb{N}^+$ esetén*

$$\mathbb{E}(I_m) = \mathbb{P}\big(X_m = \max\{X_1, \ldots, X_m\}\big) = \frac{1}{m} \tag{3.1}$$

Bizonyítás. (3.1) könnyen adódik, hiszen az abszolút folytonosság miatt

$X_1, X_2, \ldots, X_m$ között 1 valószínűséggel van szigorú maximum, az iid tulajdonság miatt pedig $X_1, X_2, \ldots, X_m$ mindegyike egyforma, azaz $\frac{1}{m}$ valószínűséggel lehet a szigorú maximum.

Az egyszerűbb jelölés kedvéért $\forall k \in \mathbb{N}^+$ esetén $H_k := H_k^{(1)} = \{X_k = X_{1,k}\}$. A függetlenség bebizonyításához azt elég lenne belátnunk, hogy $H_1, H_2, \ldots$ események függetlenek. Tehát azt kellene belátnunk, hogy $H_1, H_2, \ldots$ bármely véges kezdőszelete független. Ezt a kezdőszelet n hossza szerinti teljes indukcióval fogjuk belátni.

Az $n = 1$ eset nyilvánvaló. Tegyük fel, hogy n-re teljesül az állítás, azaz bármely n hosszú kezdőszelet független. Belátjuk, hogy ekkor $(n+1)$-re is igaz az állítás. A 3.1.3. állítás miatt $\sigma(\mathscr{F}_n)$ és H_{n+1} független. Be kéne látni, hogy $H_1, \ldots, H_{n+1}$ bármely legfeljebb $n+1$ elemű metszetének valószínűsége faktorizálódik. Ha H_{n+1}-et beválasztjuk a metszetbe, akkor a 3.1.3. állítás miatt készen vagyunk. Ha pedig H_{n+1}-et nem választjuk be a metszetbe, akkor az indukciós feltevés miatt vagyunk készen. Ezzel beláttuk a tétel állítását. □

Tehát kiderült, hogy R_n független indikátorok összegeként áll elő. Ez a tény viszonylag könnyűvé teszi az R_n-hez kapcsolódó állítások bizonyítását, hiszen független indikátorok összegére - sőt általánosságban független valószínűségi változók összegére - számos jól használható tétel ismert. A továbbiakban néhány R_n-re vonatkozó tétel és állítás kerül ismertetésre.

3.1.5. Állítás. *Legyen adott $X_1, X_2, \ldots$ független, azonos eloszlású abszolút folytonos sorozat, és $n \in \mathbb{N}_+$. Ekkor*

$$\mathbb{E}(R_n) = \sum_{k=1}^{n} \frac{1}{k} \tag{3.2}$$

$$\mathbb{D}^2(R_n) = \sum_{k=1}^{n} \left(\frac{1}{k} - \frac{1}{k^2} \right) \tag{3.3}$$

Bizonyítás. (3.1)-et felhasználva:

$$\mathbb{E}(R_n) = \mathbb{E}\left(\sum_{k=1}^{n} I_k \right) = \sum_{k=1}^{n} \mathbb{E}(I_k) = \sum_{k=1}^{n} \frac{1}{k}$$

Valamint a függetlenségből:

$$\mathbb{D}^2(R_n) = \mathbb{D}^2\left(\sum_{k=1}^{n} I_k \right) = \sum_{k=1}^{n} \mathbb{D}^2(I_k) = \sum_{k=1}^{n} \left(\frac{1}{k} - \frac{1}{k^2} \right)$$

□

Elsőre (3.2) talán meglepő eredménynek tűnik. Például $\mathbb{E}(R_{10^6}) = 14.39$. Azaz 1 millió próbálkozásból várhatóan nagyjából 14 rekord adódik. Ez valamelyest szembe megy a szemlélettel, hiszen az életben azt szokhattuk meg, hogy rekordok viszonylag gyakran adódnak. Például ha egy adott sport esetén az évenkénti legjobb eredményt vesszük a valószínűségi változóknak, akkor világos, hogy másfajta viselkedést tapasztalunk. Ebből már rögtön arra következtethetünk, hogy a továbbiakban, amikor sportrekordokat modellezünk, akkor az a feltételezés nem lesz helytálló, miszerint az évi legjobb eredmények független azonos eloszlásúak.

Viszont, ha kvantifikálni akarjuk ezt a gondolatmenetet, akkor célszerű konkrétan meghatározni R_n eloszlását. Lévén, hogy R_n független indikátorok összegére bomlik, egy adott $k \in \mathbb{N}^+, 1 \leq k \leq n$ szám esetén a $\mathbb{P}(R_n = k)$ valószínűség kiszámolása $\binom{n}{k}$ darab k tényezős szorzat összeadásával megkapható. Hiszen venni kell $\{1, \ldots, n\}$ összes k elemű részhalmazát, és a részhalmaz elemei által kijelölt helyek indikátorainak valószínűségét kell összeszorozni, majd az összes ilyen szorzatot szummázni kell. Ez a számolási módszer azonban a gyakorlatban túl költségesnek bizonyul. Ha egyszerre több valószínűségre is szükségünk van, akkor meg főleg sok ideig tartana ily módon számolni.

De szerencsére egy egyszerűbb eljárás is adható erre a problémára, amely $\mathcal{O}(n^2)$ időben egyszerre kiszámolja az összes $P(R_i = j)$ valószínűséget $i = 1, \ldots, n$, $j = 1, \ldots, i$ számokra. Az algoritmus ötlete mindössze annyi, hogy $i = 1, \ldots, n$ esetén $P(R_i = 1) = 1/i$, $P(R_i = i) = 1/i!$. Valamint ha $i \geq 3$, akkor $j = 2, \ldots, i-1$ számokra $P(R_i = j) = (1 - \frac{1}{i})P(R_{i-1} = j) + \frac{1}{i}P(R_{i-1} = j - 1)$ Így megfelelő sorrendben haladva, a kiszámolt értékeket mindig eltárolva $\mathcal{O}(n^2)$ időben megkapjuk az összes valószínűséget.

3.2. A rekordok aszimptotikus gyakorisága

A rekordok ritkaságát a következő tétel is mutatja.

3.2.1. Tétel. *Legyen adott $X_1, X_2, \ldots$ független, azonos eloszlású abszolút folytonos sorozat. Ekkor*

$$\frac{R_n}{\ln(n)} \xrightarrow{1\,vsz.} 1.$$

Bizonyítás.

Legyen $n \in \mathbb{N}^+$ esetén $I_n^* = I_n - 1/n$. Ekkor a Kolmogorov-kritérium alkalmazható I_n^*-ra, hiszen (I_n) függetlensége miatt (I_n^*) is független, valamint

$$\sum_{n=1}^{\infty} \frac{\mathbb{D}^2(I_n^*)}{n^2} \leq \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty.$$

Így tehát

$$\frac{1}{n}\sum_{k=1}^{n} I_n^* \xrightarrow{1\text{vsz.}} 0.$$

Vagyis

$$\frac{1}{n}\sum_{k=1}^{n} I_n - \frac{1}{n}\sum_{k=1}^{n} 1/k \xrightarrow{1\text{vsz.}} 0.$$

Vagyis $\frac{R_n}{n} \sim \frac{ln(n)}{n}$ 1 valószínűséggel, azaz $R_n \sim ln(n)$ 1 valószínűséggel. Ezzel beláttuk a tételt.

□

3.2.2. Tétel. *Legyen adott $X_1, X_2, \ldots$ független, azonos eloszlású abszolút folytonos sorozat. Továbbá legyen $a, b \in \mathbb{R}^+$, $a < b$. Ekkor*

$$R_{\lfloor bn \rfloor} - R_{\lfloor an \rfloor} \xrightarrow{d} Y \sim Poi\big(ln(b/a)\big).$$

Bizonyítás.

A szóban forgó valószínűségi változó felírható az eddigiekhez hasonló módon független indikátorok összegeként:

$$R_{\lfloor bn \rfloor} - R_{\lfloor an \rfloor} = \sum_{k=\lfloor an \rfloor + 1}^{\lfloor bn \rfloor} I_k.$$

Itt az indikátorok paramétereinek összege

$$\sum_{k=\lfloor an \rfloor + 1}^{\lfloor bn \rfloor} \frac{1}{k} = \sum_{k=1}^{\lfloor bn \rfloor} \frac{1}{k} - \sum_{k=1}^{\lfloor an \rfloor} \frac{1}{k} \sim ln(bn) - ln(an) = ln(b/a).$$

Így tehát a független indikátorok összegének határeloszlására vonatkozó tétel szerint

$$R_{\lfloor bn \rfloor} - R_{\lfloor an \rfloor} \xrightarrow{d} Y \sim Poi\big(ln(b/a)\big).$$

□

Habár szemmel láthatólag kevés rekordra lehet számítani, mégis $\mathbb{E}(R_n) \longrightarrow \infty$. Azonban könnyen definiálhatunk olyan rekordfogalmat, melyre a várható rekordszám már valami véges számhoz konvergál.

3.2.3. Definíció. *Legyen $X_1, X_2, \ldots$ valószínűségi változó sorozat, valamint legyen $k \in \mathbb{N}^+$ adott. Ekkor $\forall n \in \mathbb{N}^+$ esetén*

$$I_{n,k} := \prod_{j=n}^{n+k-1} I_j.$$

Tehát $I_{n,k}$ azt mutatja, hogy van-e k darab egymást követő rekord az n-edik helytől indulva. Például $k = 1$ esetben visszakapjuk a szokásos rekord fogalmat, $k = 2$ esetben ikerrekordokról beszélhetünk.

3.2.4. Definíció. *Legyen $X_1, X_2, \ldots$ valószínűségi változó sorozat, valamint legyen $k \in \mathbb{N}^+$ adott. Ekkor $\forall n \in \mathbb{N}^+$ esetén*

$$R_{n,k} := \sum_{j=1}^{n} I_{j,k}.$$

Tehát $R_{n,k}$ azt adja meg, hogy az n hosszú kezdőszeletben hány k-as rekordnak van kezdőeleme.

3.2.5. Állítás. *Legyen $X_1, X_2, \ldots$ független, azonos eloszlású, abszolút folytonos valószínűségi változók sorozata. Ekkor*

$$\mathbb{E}(R_{n,2}) \longrightarrow 1.$$

Továbbá $\exists\, R_{\infty,2}$ valószínűségi változó, melyre $R_{n,2} \overset{1\,vsz.}{\longrightarrow} R_{\infty,2}$ és $R_{\infty,2} < \infty$ 1 valószínűséggel.

Bizonyítás. A megfelelő indikátorok függetlenségét felhasználva

$$\mathbb{E}(R_{n,2}) = \mathbb{E}\left(\sum_{j=1}^{n} I_j I_{j+1}\right) = \sum_{j=1}^{n} \mathbb{E}(I_j)\mathbb{E}(I_{j+1}) = \sum_{j=1}^{n} \frac{1}{j(j+1)} \longrightarrow 1.$$

Az állítás második részéhez vegyük észre, hogy $R_{n,2}$ monoton nő, tehát (az alaphalmaz minden elemére) van határértéke. Ekkor ezen $R_{\infty,2}$ határértékre a Beppo Levi tétel miatt

$$\mathbb{E}(\lim_{n\to\infty} R_{n,2}) = \lim_{n\to\infty} \mathbb{E}(R_{n,2}) = 1.$$

Azaz $\mathbb{E}(R_{\infty,2}) < \infty$, vagyis $R_{\infty,2} < \infty$ 1 valószínűséggel.
□

3.3. Várakozási idő rekordok között

3.3.1. Definíció. *Legyen adott $X_1, X_2, \ldots$ valószínűségi változó sorozat és $n \in \mathbb{N}^+$. Ekkor*

$$K_n := \sum_{k=1}^{\infty} \prod_{j=1}^{k} (1 - I_{n+j}).$$

Tehát K_n azt adja meg, hogy X_n után mennyit kell várni a következő rekordra (pontosabban a sikertelen próbálkozások számát adja meg). Ugyanis rögzített $n, k \in \mathbb{N}^+$ esetén $\prod_{j=1}^{k}(1-I_{n+j})$ egy indikátor valószínűségi változó, mely azt mutatja, hogy volt-e legalább k darab sikertelen próbálkozás X_{n+1}-gyel kezdődően.

3.3.2. Állítás. *Legyen adott $X_1, X_2, \ldots$ független, azonos eloszlású, abszolút folytonos valószínűségi változók sorozata és $n \in \mathbb{N}^+$. Ekkor*

$$\mathbb{E}(K_n) = +\infty \tag{3.4}$$

Bizonyítás. Az állítást elég $n = 1$ esetre belátni. Elemi kombinatorikus meggondolással könnyen megkaphatjuk $\mathbb{P}(K_1 = k)$ értékét (k=0,1, ...). Ugyanis $K_1 = k$ azt jelenti, hogy $X_1, \ldots, X_{k+2}$ közül X_{k+2} a legnagyobb, $X_2, \ldots, X_{k+1}$ mindegyike pedig kisebb X_1-nél. Vagyis X_{k+2} a legnagyobb, X_1 a második legnagyobb, $X_2, \ldots, X_{k+1}$ a maradék k pozíciót pedig tetszőleges sorrendben elfoglalhatja. Ebből

$$\mathbb{P}(K_1 = k) = \frac{k!}{(k+2)!} = \frac{1}{(k+1)(k+2)}.$$

Tehát

$$\mathbb{E}(K_1) = \sum_{k=0}^{\infty}\left(k\frac{1}{(k+1)(k+2)}\right) = +\infty.$$

□

4. fejezet

Sportrekordok modellezése

Ebben a fejezetben kerülnek különféle sportrekordok modellezésre az eddigiek felhasználásával. Már minden kellő ismeret rendelkezésünkre áll ahhoz, hogy ezt a modellezést véghez vigyük. Ahogy az a bevezetésben is említést kapott, ez a fejezet bizonyos mértékben a [2] könyvben és a [7] cikkben leírtakra támaszkodik, de a konkrét számítások, adatelemzések önálló munka eredményei. A statsztikai vizsgálatokat az R programnyelv segítségével végeztem az *evd* és a *rootSolve* csomagok felhasználásával.

4.1. Síkfutás

Kétség kívül az egyik legizgalmasabb atlétikai szám a 100 valamint 200 méteres síkfutás. Usain Bolt elképesztő eredményei óta kiemelt figyelem övezi a férfi síkfutás rekordjainak alakulását. Először foglalkozzunk a 100 méteres síkfutással.

4.1.1. A tendencia javulása

Ha összehasonlítjuk egy régebbi és egy mai futam felvételeit, akkor azonnal az lehet a benyomásunk, hogy az atléták teljesítménye jelentősen javult az évek során. Ez természetesnek is tűnik, tekintve, hogy az atléták száma a Föld népességével együtt nőtt az elmúlt években. Valamint az edzési módszerek, táplálkozási ismeretek fejlődése is nagyban elősegítette a teljesítmények javulását. Ezen felül a futócipők és egyéb edzést segítő eszközök fejlődése sem elhanyagolható hatású. Emellett az is lehetséges, hogy a doppingolás is hozzájárult a teljesítmények javulásához, azonban ennek a vizsgálata nem témája a dolgozatnak.

A tendencia javulásával kapcsolatos sejtésünket, észrevételünket fogalmazzuk meg kicsit precízebben. Legyen adott m év évi legjobb eredménye (-1)-gyel szorozva: $y_1, y_2, \ldots, y_m$. Ekkor a (1.1) érvelés miatt észszerű ezekre valamilyen

$Y_1, Y_2, \ldots, Y_m$ extrémérték-eloszlásokból származtatott értékekként tekinteni. Azt rögtön sejtjük, hogy a teljesítmények folyamatos javulásából adódóan $Y_1, \ldots, Y_m$ nem azonos eloszlásúak. Ebből kifolyólag a fejezet fő célja ezen változás, fejlődés modellezése lesz.

A továbbiakban feltételezzük, hogy $Y_1, \ldots, Y_m$ függetlenek. Ennek a feltételezésnek a jogossága persze vitatható. A valóságban itt tényleges függetlenségről nem beszélhetünk, de a modellalkotást hatalmas mértékben könnyíti ez a feltételezés, és reményeink szerint többet nyerünk a modellünk egyszerűsödésével, mint amennyit a valóság effajta torzításával veszítünk. Később, amikor már elkészült a modellünk, akkor vizsgálhatjuk a reziduálisok függetlenségét.

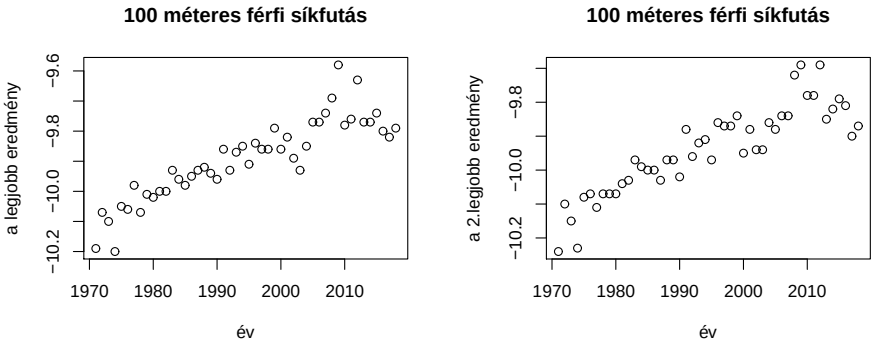
Ha több adat rendelkezésünkre állna, akkor tudnánk tömbösíteni az éveket, és ekkor a függetlenség még jogosabb feltételezés lenne. De sajnos ezen a téren nincs könnyű dolgunk, hiszen gépi időmérésű eredmények csak körülbelül az elmúlt 50 évből vannak. A kézi időmérés pontatlansága ilyen rövid távon pedig nagyban hátráltatná a modellalkotást.

Vagyis $i = 1, 2, \ldots, m$ esetén $Y_i \sim GEV(\mu_i, \sigma_i, \xi_i)$. A feladat tehát a (μ_i, σ_i, ξ_i) paraméterek meghatározása. Ebben a kontextusban az előbbi, köznapi nyelven megfogalmazott észrevétel tehát arra utal, hogy a paramétereknek változnia kell, méghozzá oly módon, hogy az Y_i valószínűségi változókból várhatóan jobb eredmények származzanak. Például elvárható, hogy $1 \leq i < j \leq m$ esetén $F_{Y_i} \geq F_{Y_j}$ teljesüljön. Sőt elvárjuk a szigorú egyenlőtlenség teljesülését is az olyan helyeken, ahol egyik eloszlásfüggvény sem 0 vagy 1.

Mielőtt elkezdenénk tárgyalni, hogy melyik paraméter hogyan változik, érdemes leellenőrizni, hogy valóban szükség van-e a paraméterek változtatására a modellben. Tekintsük 1971-től 2018-ig a 100 méteres férfi síkfutás éves legjobb eredményeit (-1)-gyel szorozva.

$$\left(y_i^{(1)}\right)_{i=1}^{48} = -(10.19, 10.07, 10.10, 10.20, 10.05, 10.06, 9.98, 10.07, 10.01, 10.02,$$
$$10.00, 10.00, 9.93, 9.96, 9.98, 9.959.93, 9.92, 9.94, 9.96, 9.86, 9.93, 9.87, 9.85, 9.91,$$
$$9.84, 9.86, 9.86, 9.79, 9.86, 9.82, 9.89, 9.93, 9.85, 9.77, 9.77, 9.74, 9.69, 9.58, 9.78,$$
$$9.76, 9.63, 9.77, 9.77, 9.74, 9.80, 9.82, 9.79)$$

Itt az előző fejezet jelöléseivel élve a rekordok és minimumrekordok számaira $R_{48} = 14$, $R_{48}^{(-1)} = 2$. Független, azonos eloszlású $Y_1, \ldots, Y_{48}$ esetén $\mathbb{E}(R_{48}) = \mathbb{E}(R_{48}^{(-1)}) = 4,46$. Továbbá az előző fejezetben leírt algoritmussal kiszámolható, hogy $\mathbb{P}(R_{48} \geq 14) = 5,98 \cdot 10^{-6}$ adódik független, azonos eloszlású esetben. Ez a kis valószínűség aligha meglepő az eddigiek alapján, hiszen láttuk az előző fejezetben, hogy 1 millió próbálkozásból adódik várható értékben 14,39 rekord.



(a) A legjobb idők

(b) A 2. legjobb idők

1. ábra. Az évenkénti első és második legjobb idők 1971-től 2018-ig

Szóval a 14 rekord elérése 48 próbálkozásból valóban rendkívül valószerűtlennek tűnik.

Világos tehát, hogy $Y_1, \ldots, Y_n$ nem azonos eloszlásúak, vagyis javuló tendencia figyelhető meg a paraméterek alakulásából adódóan. Ezt a javuló tendenciát az 1. ábra is szemlélteti, melyen az évenkénti legjobb és második legjobb eredmények vannak ábrázolva. Fontos, hogy ezen az ábrán és a továbbiakban mindig egy adott évben egy atlétának csak a legjobb eredményét vesszük figyelembe. Például 2013-ban Usain Bolt futotta az első három legjobb időt, tehát abban az évben, amit mi most 2. legjobb eredménynek veszünk, az összességében csak a 4. legjobb volt. Ezáltal egyrészt elkerülhetjük, hogy az egyéni teljesítmények túl erősen befolyásolják a modellezést. Másrészt a GEV modell használata akkor indokolt, ha egy adott év legjobb eredményei sok független valószínűségi változó maximumából vagy minimumából adódnak. Ha nem tömbösítünk, akkor ez a megkívánt függetlenség nem teljesül.

4.1.2. A paraméterek becslése

Az eddigiek alapján tehát $Y_i \sim GEV(\mu_i, \sigma_i, \xi_i)$ és feladatunk μ_i, σ_i, ξ_i meghatározása . A [2] könyv és [7] cikk szerint minden sportot kontans σ és ξ paraméterekkel, valamint folytonosan változó μ_i paraméterrel érdemes modellezni. Egészen konkrétan síkfutás esetén $\mu_i = \mu^{(1)} - \mu^{(2)} \exp(-\mu^{(3)} \cdot i)$ alakban érdemes a helyparamétert keresni, megfelelő $\mu_1, \mu_2, \mu_3 \in \mathbb{R}$ paraméterekkel. A munkánkat valóban nagyban megkönnyítené, ha az alakparamétert és a skálaparamétert konstansnak vehetnénk. Azonban ezen egyszerűsítés jogossága némi indoklást igényel.

Tekintsünk néhány évet, melyről elegendő mennyiségű adatunk van. A [8]

weboldalon megtalálhatóak az éves legjobb idők. Tipikusan a korábbi évekből kevesebb adat maradt fent, azonban az elmúlt 15-20 évben évenként többszáz, vagy akár többezer atlétának is fel van jegyezve a legjobb eredménye. Például vegyük a 2005, 2009, 2014 és 2018 éveket. Itt rendre a legjobb kb 600, 2600, 5000 és 5500 atléta eredménye található meg. Az eddigi jelölésekkel $i = 35, 39, 44, 48$ esetén rendelkezünk Y_i maximum vonzási tartományában lévő valószínűségi változókból álló nagy méretű rendezett mintával.

Ennek fényében rögtön az juthat eszünkbe, hogy a Hill-becslés jól használható ebben a helyzetben ξ_i becslésére. Azonban, az 1. fejezetben látott érvelés szerint negatív alakparaméterre számíthatunk, a Hill-becslés pedig csak $\xi > 0$ esetben használható. Ez a probléma orvosolható, ha a hagyományos Hill-becslés helyett annak a Dekkers–Einmahl–de Haan-féle változatát tekintjük, mely a Hill-becslés egy kiterjesztése, olyan értelemben, hogy tetszőleges $\xi \in \mathbb{R}$ esetén alkalmazható.

A módszer $\forall k = 1, \ldots, n$ esetén szolgáltat egy $\hat{\xi}(k)$ becslést az alakparaméterre, ahol k a használt legnagyobb mintaelemek száma, n pedig az összes mintaelem száma. Ha a felhasznált minta valamilyen extrémérték-eloszlás maximum vonzási tartományában van, és valamilyen $\delta \in \mathbb{R}_0^+$ számra $k/(\ln n)^\delta \to \infty, n \to \infty$, akkor erősen konzisztens a becslés. Ez és a becslés egyéb tulajdonságai bővebben a [9] cikkben kerülnek tárgyalásra. A formula a tehát következő.

$$\hat{\xi}(k) = 1 + H_n^{(1)}(k) + \frac{\left(H_n^{(1)}(k)\right)^2}{2H_n^{(2)}(k)},$$

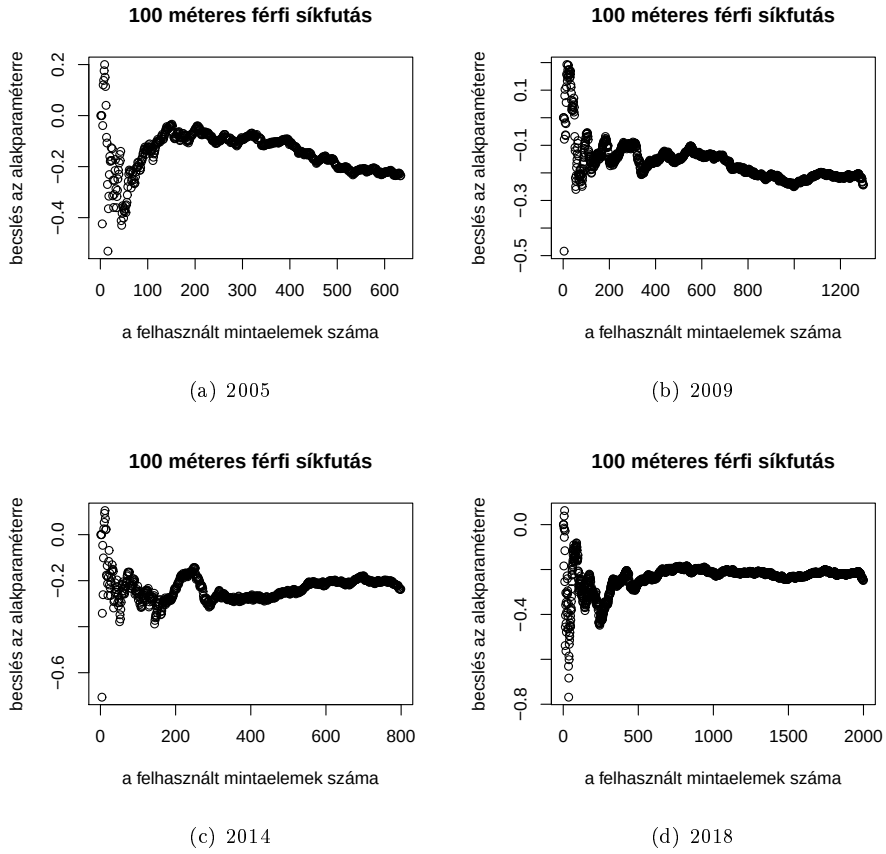
ahol

$$H_n^{(1)}(k) = \frac{1}{k}\sum_{j=1}^{k}(\ln X_{j,n} - \ln X_{k+1,n}),$$

$$H_n^{(2)}(k) = \frac{1}{k}\sum_{j=1}^{k}(\ln X_{j,n} - \ln X_{k+1,n})^2.$$

A 100 méteres síkfutási eredmények csak 2 tizedesjegy pontosságig kerülnek rögzítésre. Ekkor lényegében diszkrét valószínűségi változóként viselkednek az eredményeket szolgáltató valószínűségi változók, márpedig a diszkrét valószínűségi változók tipikusan nem tartoznak egy extrémérték-eloszlás maximum vonzási tartományába sem. Tehát a becslések kiszámítása előtt szükséges a mérési eredmények simábbá tétele. Ezt elérhetjük oly módon, hogy véletlenszerűen egy $-0,005$ és $0,005$ közti számot adunk a mérési eredményekhez, majd rendezzük a mintát. Így már a szóban forgó valószínűségi változók tekinthetők folytonosnak, és a módszer elvégezhető.

A 2. ábrán a kiválasztott 4 évben az alakparaméterre adott becslések láthatóak k függvényében. Az ilyen ábrák értelmezése nem mindig könnyű feladat,



(a) 2005

(b) 2009

(c) 2014

(d) 2018

2. ábra. ξ Dekkers–Einmahl–de Haan becslése a használt mintaelemek számának függvényében

de az általánosan elfogadott alapelv az, hogy a grafikonnak egy vízszinteshez közeli részét kell keresni a megfelelő becslés leolvasásához. A 2. ábra alapján azt mondhatjuk, hogy az alakparaméter nagyjából változatlannak tekinthető és a $\hat{\xi} = -0.2$ egy észszerű becslés.

A korábbi évek alakparamétereire ugyan nem láttunk becslést, mégis jogosnak tűnik az a feltételezés, hogy ha a 2005-től 2018-ig terjedő intervallumban nem változik az alakparaméter, akkor az 1971-től 2005-ig terjedő időszakban sem. Egyelőre tehát tekintsük az alakparamétert konstansnak. Valamint tekintsük a skálaparamétert is konstansnak. Majd mindkét esetben, miután a konstans értéket feltételező modellt megalkottuk, azt fogjuk tapasztalni, hogy a paraméterek változásának megengedése nem eredményez olyan mértékű javulást, amely indokolná a paraméterek számának növelését.

Még a helyparaméter alakulását kell meggondolnunk. Egyrészt az eddigiekből következően elvárjuk, hogy μ_i monoton növekedjen. Másrészt μ_i korlátossága is jogos elvárás, hiszen rögzített σ és $\xi < 0$ mellett μ_i korlátlan növekedése azt jelentené, hogy $\mathbb{P}(Y_i > 0) \overset{i\to\infty}{\longrightarrow} 1$.
Megjegyzendő, hogy nem konstans skálaparaméter és alakparaméter mellett ezeknek a tulajdonságoknak nem feltétlenül kellene eleget tennie μ_i-nek, elképzelhető lenne valamilyen bonyolultabb összefüggés a paraméterek között, mely ugyanúgy teljesítené a kívánt feltételeket. Azonban a modellalkotás szempontjából sokkal szerencsésebb az egyszerűbb, monotonitást kihasználó gondolatmenet. Ezek az érvek tehát a konstans σ és ξ értéket használó modell mellett szólnak. De erre a kérdésre még a modellalkotás után visszatérünk.

Ezen megfontolások alapján valóban ideális jelöltnek tűnik a $\mu_i = \mu^{(1)} - \mu^{(2)} \exp(-\mu^{(3)} \cdot i)$ alakú paraméterezés. Vagyis összesen 5 paraméterrel dolgozunk a továbbiakban. Mondhatnánk azt, hogy ξ ne legyen paraméter, hanem rögzítsük a Dekkers–Einmahl–de Haan becslés által kapott értékre. Azonban jó ellenőrzési alapul szolgál, ha ξ-re kétféle becslést is látunk. A paraméterekre maximum likelihood becslést fogunk alkalmazni. A maximum likelihood becslést a már említett 1971-től 2018-ig terjedő mintával végezzük. Minden évben tekintjük az első 10 legjobb eredményt, illetve 10-nél kisebb elemszámú éves minta esetén a feljegyzett összes legjobb eredményt. Ekkor tehát adott egy 48 elemű minta, mely független valószínűségi vektorváltozókból adódott. Így a likelihood egy 48 tényezős szorzat lesz, ahol az egyes tényezőket 2.2.5. tétel segítségével kaphatjuk meg.

Legyen tehát az i. év j. legjobb eredménye (-1)-gyel szorozva $y_i^{(j)}$. Ekkor feltéve, hogy a 2.2.5. tételben leírt feltételek teljesülnek, a likelihood függvényre

a következő aszimptotikus közelítés adható.

$$f(y) = \prod_{i=1}^{48} \left(H(y_i^{(r_i)}) \prod_{j=1}^{r_i} \left(\frac{h_i(y_i^{(j)})}{H_i(y_i^{(j)})} \right) \mathbb{1}_{D(k)}(y_i) \right),$$

ahol r_i az i. éves rendezett mintának a becslésben felhasznált elemszáma, h_i és H_i pedig rendre $GEV(\mu_i, \sigma, \xi)$ sűrűségfüggvény és eloszlásfüggvény. Ebből tehát a megkapható a loglikelihood, az indikátorokat el is hagyhatjuk, a minta úgyis jól van rendezve.

$$\begin{aligned} l(y) &= \sum_{i=1}^{48} \left(\ln H(y_i^{(r_i)}) + \sum_{j=1}^{r_i} \left(\ln h_i(y_i^{(j)}) - \ln H_i(y_i^{(j)}) \right) \right) = \\ &= -\sum_{i=1}^{48} \left(\left(1 + \xi \frac{y_i^{(r_i)} - (\mu^{(1)} - \mu^{(2)} e^{\mu^{(3)} \cdot i})}{\sigma} \right)^{-1/\xi} + r_i \ln \sigma + \right. \\ &\left. + \sum_{j=1}^{r_i} \left((1 + 1/\xi) \ln \left(1 + \xi \frac{y_i^{(j)} - (\mu^{(1)} - \mu^{(2)} e^{\mu^{(3)} \cdot i})}{\sigma} \right) \right) \right) \end{aligned},$$

Ennek a függvénynek a maximalizálása analitikus módon nyilván reménytelen, így tehát numerikus módszerekhez kell folyamodnunk. Szerencsére az R nyelv rendelkezik a numerikus maximalizáláshoz szükséges függvényekkel. Először is inputként be kell adnunk az R-nek a megfelelő időket. Ezt egy mátrix formájában érdemes megtenni. Ezután az a teendőnk, hogy definiáljuk R-ben az előbb látható l függvény (-1)-szeresét. A negatív előjel azért szükséges, mert az optimalizálásra megírt függvények nagy része minimumkeresésre van beállítva. Ezután valamelyik ilyen optimalizáló függvénynek be kell adnunk a $-l$ függvényt, valamint 5 számot, melyből az optimalizálást indítja a program. A kiindulási paraméterek meghatározása nem feltétlenül nyilvánvaló, és döntően tudja befolyásolni a módszer hatékonyságát. Ha szerencsésen választjuk meg a kezdő paramétereket, akkor a módszer által meghatározott lokális minimum egyben globális minimum is. Ezen felül magának a módszernek a választása sem triviális, különböző módszerek különböző eredménnyel szolgálhatnak.

Érdemes kísérletezni különböző R-beli függvényekkel és különböző kiindulási értékekkel. Mindenképp olyan kiindulási értéket kell megadunk, melyre kiértékelhető a loglikelihood függvény, ellenkező esetben egyből hibát jelez az R, és leáll az optimalizáló algoritmus. Először kezdjünk konstans helyparaméterrel, azaz $\mu^{(2)} = \mu^{(3)} = 0$ értéket vegyünk, ugyanis nem könnyű elsőre látni, hogy a változást leíró paraméterek hogyan befolyásolják a loglikelihood értékét. Legyen mondjuk $\mu^{(1)} = -10$. Az alakparamétert indítsuk a Dekkers-Einmahl-de Haan-féle becslésből kapott értékről, azaz $\xi = -0,2$ legyen a kezdőértéke. Elsőre a skálaparamétert sem könnyű megtippelni, induljunk tehát a semlegesnek tűnő

$\sigma = 1$ értékről.

Nézzük meg kettő R-beli optimalizáló függvénnyel is, hogy milyen eredményt kapunk ezekkel a kiindulási értékekkel. A továbbiakban közölt loglikelihood értékek a (-1)-gyel való visszaszorzás után értendők. Az *optim* nevű függvény $1370,696$ értéket ad. Az *nlm* nevű függvénnyel $1354,854$ értéket kapunk. Nézzük meg mi történik, ha mondjuk $\mu^{(1)} = -9,5$ értékre változtatjuk a kiindulási helyparamétert. Ekkor az *optim* függvény által adott érték lecsökken $1343,778$-ra, viszont az *nlm* által adott érték megnő $1378,726$-ra. Módosíthatjuk a változást leíró kiindulási paramétereket is. Azt tapasztaljuk, hogy az *optim* függvény kisebb változtatásokra is nagy ingadozást mutat, de egyszer sem ad 1378-nál jobb értéket, míg az *nlm* legtöbbször az $1378,726$ értéket adja. Így esetünkben jobban használhatónak tűnik az *nlm* függvény, és azt tapasztaljuk, hogy $\mu^{(1)} = -9,5$ kiindulási értékre és tetszőleges olyan többi kezdőparaméterekre, melyre kiértékelhető a loglikelihood az 1378.726 érték adódik. Eltérő $\mu^{(1)}$ kiindulási értékekre nem ilyen stabil viselkedést mutat az *nlm* függvény, és kisebb loglikelihood értéket is ad. Tehát úgy tűnik, hogy megtaláltuk a maximális loglikelihoodot.

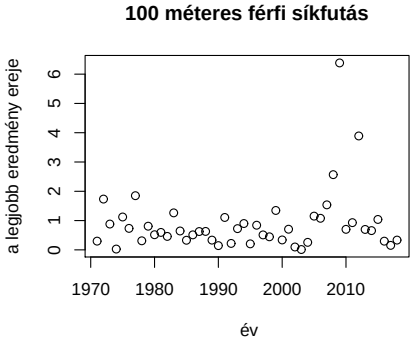
Maxmimumhelynek $\hat{\mu}^{(1)} = -9,7159, \hat{\mu}^{(2)} = 0,4760, \hat{\mu}^{(3)} = 0,04299, \hat{\sigma} = 0,05789$, $\hat{\xi} = -0,1709$ adódik. Nincs módunk megbizonyosodni matematikai precízséggel arról, hogy ez valóban globális maximumhely, de az előbbi leírt kísérletezés alapján ez bizonyult a legjobbnak. Mindenesetre bíztató, hogy az alakparaméterre a két látott becslés egymáshoz közeli értéket adott.

4.1.3. Usain Bolt eredményei

Először is nézzük meg azt, hogy mennyire mondható erős rekordnak Usain Bolt eredménye, azaz abban az évben, amelyben a rekordot beállító eredménye született, mekkora volt a valószínűsége egy legalább olyan jó idő elérésének. Legyen $p_i = \mathbb{P}(Y_i \geq -9,58)$. A rekordot beállító eredmény 2009-ben, azaz a 39. évben született, vagyis a keresett valószínűség

$$p_{39} = 1 - \exp\left(- \left(1 + \hat{\xi}\frac{-9,58 - (\hat{\mu}^{(1)} - \hat{\mu}^{(2)}e^{-\hat{\mu}^{(3)}\cdot 39}}{\hat{\sigma}}\right)^{-1/\hat{\xi}}\right) = 0,0017.$$

Ez egy rendkívül erősnek mondható eredmény. Ha nem javulna a helyparaméter, akkor ez jóformán megdönthetetlen rekordnak bizonyulna. Ez a valószínűség kvantifikálja valamilyen szinten azt az érzést, hogy Usain Bolt addig sosem látott fölénnyel utasította maga mögé a mezőnyt. Ezt a fölényt a 3. ábra is mutatja, melyen a $-\ln \mathbb{P}\big(Y_i \geq y_i^{(1)}\big)$ értékek vannak ábrázolva, azaz minden évhez rendeltünk egy mérőszámot, mely bizonyos értelemben az abban az évben született legjobb eredmény erősségét mutatja. Az ábrán látható 3 kiugró érték mind Usain Bolt nevéhez fűződik. Az a tény, hogy Usain Bolt ennyire



3. ábra. $-\ln \mathbb{P}\big(Y_i \geq y_i^{(1)}\big)$ értéke i függvényében

kiugró eredményeket produkált felveti a jogosságát egy olyan modellnek, amely őt külön kezeli. Azaz a valóságot talán jobban tükrözi, ha úgy járunk el, hogy Usain Bolt eredményeit egyszerűen kivesszük a mintából, és elfogadjuk, hogy az ő eredményei valamilyen más eloszlásból származtak. Ennek a bizonyos eloszlásnak a becslése most minket nem érdekel, hiszen Usain Bolt már visszavonult, így a rekordok jövőbeli alakulásának szempontjából ez az eloszlás nem releváns.

Ezzel a módosítással jelentős mértékben megváltoznak a becsült paraméterek. Egészen pontosan $\hat{\mu}^{(1)} = -9,7450$, $\hat{\mu}^{(2)} = 0,4554$, $\hat{\mu}^{(3)} = 0,04668, \hat{\sigma} = 0,0493$, $\hat{\xi} = -0,2635$ adódik. A két modellből kapott eredmények különbségét a 4. ábra is mutatja, melyen az egyes évekhez tartozó két becsült eloszlás mediánja, 75%-os kvantilise és 95%-os kvantilise látható.

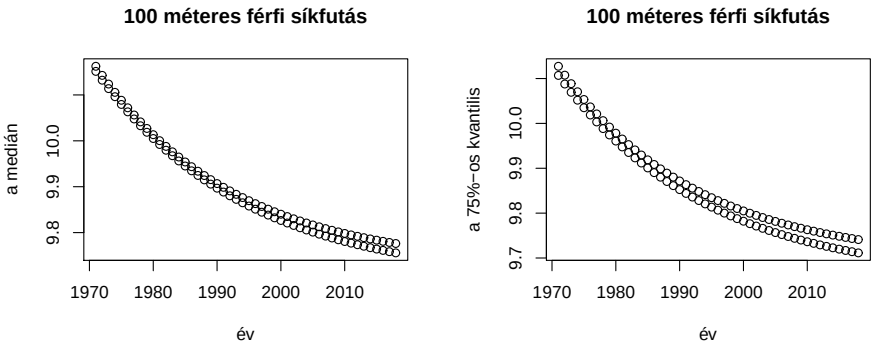
Az új paraméterekkel is megnézhetjük az egyes rekordok erősségeit. Azt tapasztalhatjuk, hogy itt nem adódik elfogadhatatlanul erős rekord. Konkrétan $\max\big\{\mathbb{P}(Y_i \leq y_i) : 1 \leq i \leq 48\big\} = 0,0119$. Így tehát jobbnak tűnik a Usain Bolt eredményeit figyelmen kívül hagyó modell.

4.1.4. A modell ellenőrzése

Láttuk tehát, hogy Usain Bolt eredményeinek figyelmen kívül hagyásával jobb modell adódik. De érdemes ennek a modellnek is ellenőrizni a helyességét. Először érdemes lehet megnézni, hogy jogos-e a konstans σ és ξ érték feltételezése.

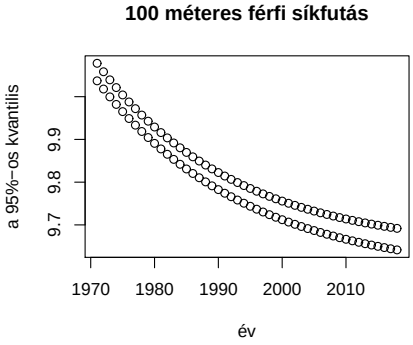
Nézzük meg, hogyan alakul a becslés, ha a skálaparaméter minden évben tetszőleges lehet. A becslés eredménye a 5. ábrán látható. Azt tapasztaljuk, hogy főként $0,04$ és $0,06$ közötti értékek adódnak. Ezen felül nem igazán figyelhető meg semmilyen tendencia sem. Illetve megjegyzendő, hogy az értékek átlaga 0,0514, azaz közel esik a konstans skálaparamétert feltételező becsléshez.

A loglikelihood értéke nagyjából 23-mal nőtt. Ez a növekedés nem teszi in-



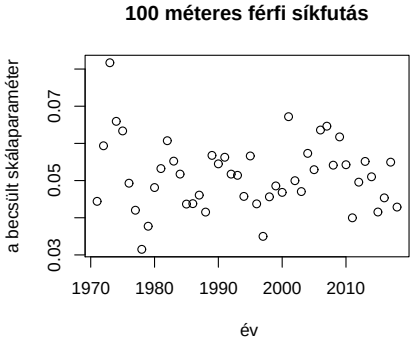
(a) medián

(b) 75%-os kvantilis



(c) 95%-os kvantilis

4. ábra. Az egyes években a két becsült eloszlás mediánja, 75%-os és 95%-os kvantilise (-1)-gyel visszaszorozva, ahol a felső értékek tartoznak a Usain Bolt eredményeit figyelmen kívül hagyó modellhez



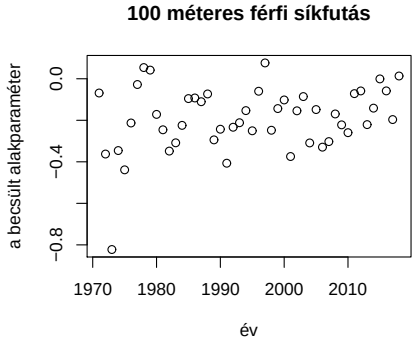
5. ábra. Maximum likelihood becslés, ha a skálaparaméter tetszőlegesen változhat

dokolttá a 47 új paraméter bevezetését.
Érdemes egy likelihood hányados próbát elvégezni. A nullhipotézis (H_0) legyen az, hogy $\sigma_i \equiv \hat{\sigma} = 0,0493$, az ellenhipotézis ($H_1$) pedig legyen az, hogy $i = 1, \ldots, 48$ esetén σ_i az előbbi, 47 paraméterrel többet használó modell becslésével egyezik meg. A többi paramétert H_0 és H_1 esetén is a megfelelő likelihood becslés adja meg.

Ismert, hogy a likelihood hányados logaritmusának kétszerese aszimptotikusan χ^2_s eloszlású, ahol s az új paraméterek száma, jelen esetben 47. Ez azonban aszimptotikus próba lévén nagy elemszámú mintát igényel. Tehát jelen esetben nem biztos, hogy jól alkalmazható ez az aszimptotikus próba. A próbát elvégezve $p = 0,51$ adódik. Ez H_0 elfogadása mellett szól, de a mintaelemek alacsony száma miatt a módszer pontossága megkérdőjelezhető.

Ha nem akarjuk az aszimptotikus eredményt használni, akkor szimulációs módszerekhez folyamodhatunk. Elsőre az az ötlet adódhat, hogy generáljunk mind a 48 évben egy vektort H_0-t feltételezve a 2.2.5. tételben szereplő eloszlással. Majd ezen véletlen vektorokra számoljuk ki a H_1 melletti és H_0 melletti megfelelő sűrűségfüggvények helyettesítési értékét. Ezután szorozzuk össze az egyes években kapott értékeket, majd osszuk el az így kapott H_1 melletti szimulált likelihoodot a H_0 melletti szimulált likelihooddal. Ha ezt a szimulációt sokszor elvégezzük, akkor ahányadrészében az eseteknek átlépte a szimulált likelihood hányados a tényleges likelihood hányadost, körülbelül annyi lesz a próba p értéke. A probléma az, hogy ezeknek a véletlen vektoroknak a generálása meglehetősen bonyolult. Így meg kell elégednünk annyival, hogy csak az évenkénti első legjobb eredményt vesszük figyelembe. Tehát az egyes éveknek megfelelő GEV eloszlásokkal generálunk véletlen számokat. Így el tudjuk végezni a szimulációt sokszor, hiszen véletlen számok generálása GEV eloszlással viszonylag gyorsan és könnyen elvégezhető R-ben. Az említett szimulációt 10^6-szor elvégezve $p = 0,12$ adódik. Ez alapján a nullhipotézis még elfogadhatónak tűnik.

Ennyi paraméter bevezetése nyilván a legszélsőségesebb eset volt. Lehet próbálkozni kevesebb paraméterrel is. De láthatóan a likelihood függvényt maximalizáló σ_i értékek nem követnek semmilyen nevezetesebb görbét, így arra számítunk, hogy nem fognak jelentős javulást eredményezni a kevesebb paramétert használó modellek sem. Például, ha $\sigma_i = \sigma^{(1)} \cdot i + \sigma^{(0)}$ alakban keressük a skálaparamétert, akkor nem javul szinte egyáltalán a becslés, és 0-hoz közeli $\hat{\sigma}^{(1)}$ értéket kapunk. Ha $\sigma_i = \sigma^{(1)} \mathbb{1}_{[0,24]} + \sigma^{(2)} \mathbb{1}_{[25,48]}$ alakban keressük a skálaparamétert, akkor $0,16$-tal nő a loglikelihood. Ekkor a likelihood hányados próbát hasonlóan elvégezve a szimulációs módszerrel körülbelül $p = 0,5$ adódik, az aszimptotikus eredményből pedig $p = 0,6$ adódik. Egyéb esetek is vizsgálhatóak, ezek részletes tárgyalásától most eltekintünk. Valamint attól is, hogy indokolt lehet-e a nem konstans skálaparaméter feltételezése, abban az esetben, ha μ_i-t valami más



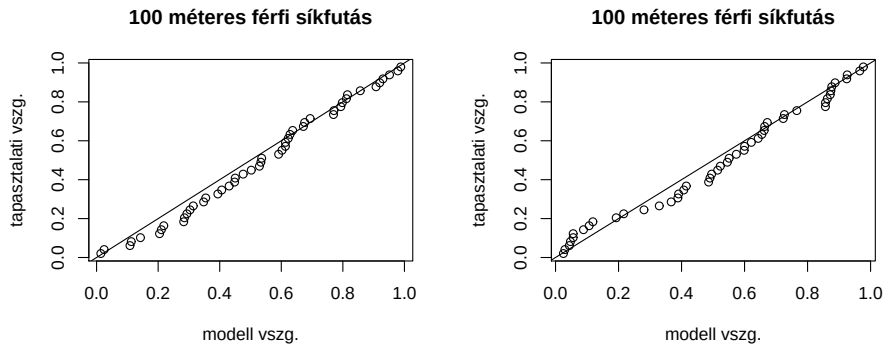
6. ábra. Maximum likelihood becslés, ha az alakparaméter tetszőlegesen változhat

függvény írja le. Összességében tehát úgy látjuk, hogy konstansnak tekinthető a skálaparaméter.

Most nézzük meg mi történik, ha az alakparaméter minden évben tetszőleges lehet. A becslés eredménye a 6. ábrán látható. Hasonlóakat tapasztalunk, mint a skálaparaméter esetében. A loglikelihood értéke körülbelül 22-vel nőtt. Ez szintén nem teszi indokolttá a 47 új paraméter bevezetését. A skálaparaméter tárgyalásánál látottakhoz hasonlóan most is elvégezhető a likelihood hányados próba és az aszimptotikus próba is. Itt 10^6-szor elvégezve a szimulációt körülbelül $p = 0,34$ adódik, az aszimptotikus próba pedig $p = 0,6$ értéket ad. Ez alapján H_0 most is elfogadhatónak tűnik. Itt is meggondolható, hogy kevesebb új paraméter bevezetése esetén mi történik, illetve más μ_i függvény esetén hogyan alakul a becslés. Tehát azt mondhatjuk, hogy az alakparaméter is konstansnak tekinthető.

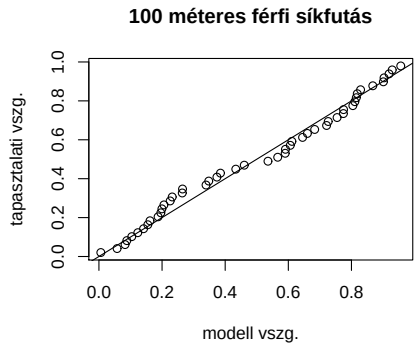
Érdemes egyéb módszerekkel is ellenőrizni a becslést. Mivel $i = 1, \ldots, 48$-ra $Y_i \sim GEV(\mu_i, \sigma, \xi)$, így $Y_i - \mu_i \sim GEV(0, \sigma, \xi)$. Vagyis ha kivonjuk az egyes évek legjobb eredményéből az adott év alakparaméterét, akkor azonos eloszlású mintát alkotnak a legjobb eredmények. Hasonlóan adódik, hogy $k = 1, \ldots, 10$ esetén, ha a k. legjobb eredményekből is kivonjuk a helyparamétert, akkor azonos eloszlású mintát kapunk, méghozzá $H^{(k)}$ eloszlásfüggvénnyel, ahol H $GEV(0, \sigma, \xi)$ eloszlásfüggvény és $H^{(k)}$ a 2.2.1. definícióban szereplő eloszlásfüggvény.

Legyen $k = 1, \ldots, 10$ -re a k. legjobb eredményekből centrálással és rendezéssel kapott minta $z_1(k) < \ldots < z_{n_k}(k)$ ahol n_k a k. legjobb eredményekből kapott minta hossza (ez nem minden esetben 48, valahol hiányoznak eredmények). Erre a mintára elkészíthetjük a megfelelő $P - P$ és $Q - Q$ plotokat. A $P - P$ plot $k = 1, \ldots, 10$ esetén a $\left\{ \left(H^{(k)}\big(z_i(k)\big), \frac{i}{n_k+1} \right) : i = 1, \ldots, n_k \right\}$ pon-

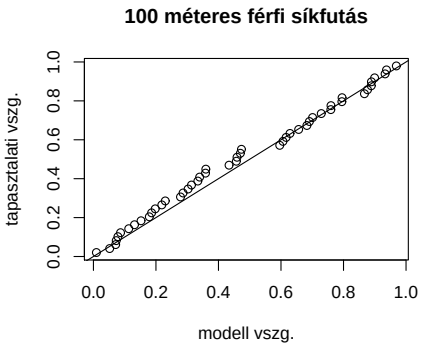


(a) $k = 1$

(b) $k = 2$



(c) $k = 3$



(d) $k = 4$

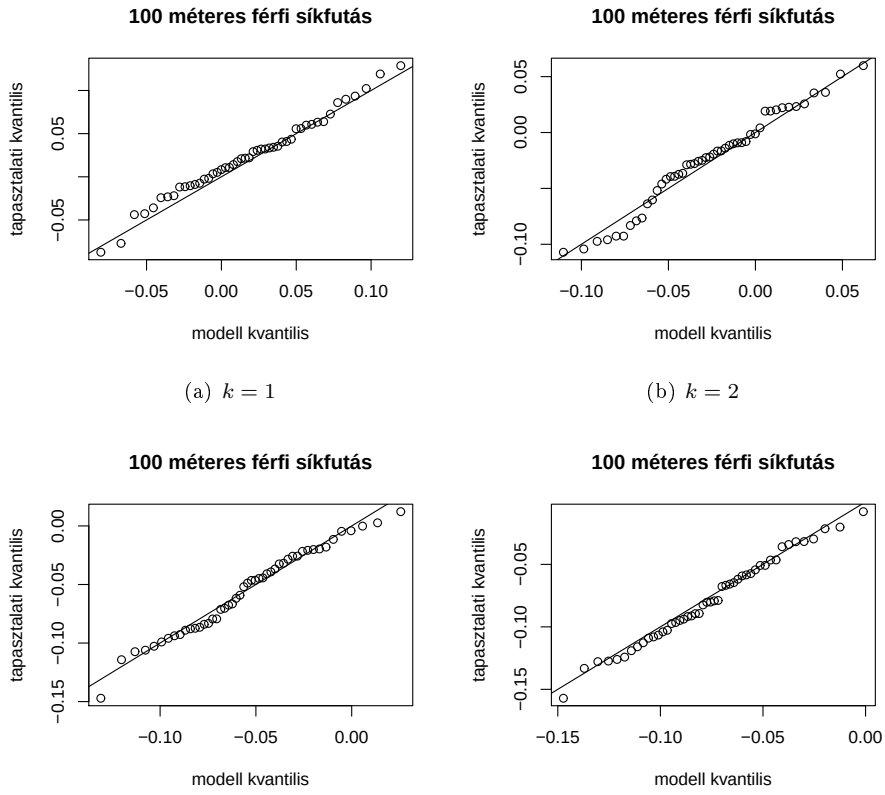
7. ábra. $P - P$ plot a k. legjobb eredményekre

tokból áll. A $Q - Q$ plot pedig a $\left\{\left(\left(H^{(k)}\right)^{-1}\left(\frac{i}{n_k+1}\right), z_i(k)\right) : i = 1, \ldots, n_k\right\}$ pontokból. A $Q - Q$ plot elkészítésénél, ha $k \geq 2$, akkor a kvantilis függvény csak numerikus invertálással kapható meg. A $P - P$ plot a 7. ábrán, a $Q - Q$ plot pedig a 8. ábrán látható. Ezek alapján azt mondhatjuk, hogy az illeszkedés megfelelőnek tűnik.

4.1.5. Rekordok a jövőben

Feltételezzük, hogy a jövőben is jól fogja jellemezni az eredmények alakulását a megalkotott modell. Így becslést adhatunk a rekordok jövőbeli alakulására vonatkozóan. Először nézzük meg, hogy várható értékben mikor fog megdőlni Usain Bolt rekordja. Mielőtt számolni kezdenénk érdemes figyelembe venni, hogy a 65. év (2035) előtt 0 a valószínűsége a rekord megdöntésének, hiszen akkor még a megfelelő eloszlások jobb végpontja is kisebb $(-9,58)$-nál.

Legyen K az a valószínűségi változó, mely megadja, hogy 2035-tól számít-



(a) $k = 1$ (b) $k = 2$

(c) $k = 3$ (d) $k = 4$

8. ábra. $Q - Q$ plot a k. legjobb eredményekre, $GEV(0, \hat{\sigma}, \hat{\xi})$ eloszlásra való transzformálással

va hányadik évben dől meg Usain Bolt rekordja. Ugyan $\mathbb{E}(K)$ pontos értékét nem tudjuk meghatározni, könnyen tudjuk közelíteni tetszőleges pontosságig. A közelítés ötlete a következő. Vegyünk valamilyen $N \in \mathbb{N}, N \geq 65$ számot, és tekintsük a következő összefüggést.

$$\mathbb{E}(K) = \sum_{n=1}^{\infty} \mathbb{P}(K \geq n) = 1 + \sum_{n=65}^{N} \left(\prod_{i=65}^{n} (1-p_i) \right) + \sum_{n=N+1}^{\infty} \left(\prod_{i=65}^{n} (1-p_i) \right) =$$

$$=1 + \sum_{n=65}^{N} \left(\prod_{i=65}^{n} (1-p_i) \right) + \prod_{i=65}^{N} (1-p_i) \sum_{n=N+1}^{\infty} \left(\prod_{i=N+1}^{n} (1-p_i) \right).$$

Ebben a kifejezésben $\sum_{n=N+1}^{\infty} \left(\prod_{i=N+1}^{n} (1-p_i) \right)$ felülről becsülhető $\sum_{n=1}^{\infty} (1-p_{65})^n = \frac{1}{p_{65}} - 1 < 10^{13}$ értékkel.

Tehát, ha $\mathbb{E}(K)$-t $\hat{\mathbb{E}}_N(K) = 1 + \sum_{n=65}^{N} \left(\prod_{i=65}^{n} (1-p_i) \right)$ értékkel becsüljük, akkor

$$\hat{\mathbb{E}}_N(K) - \mathbb{E}(K) = \prod_{i=65}^{N} (1-p_i) \sum_{n=N+1}^{\infty} \left(\prod_{i=N+1}^{n} (1-p_i) \right) < 10^{13} \prod_{i=65}^{N} (1-p_i).$$

Például $N = 2 \cdot 10^5$ választással $\prod_{i=65}^{N} (1-p_i) = 6,60 \cdot 10^{-27}$, azaz a hiba kisebb, mint $6,6 \cdot 10^{-14}$. Maga a becslés könnyen elvégezhető, a megfelelő szorzatokat kell rekurzívan kiszámolni és összeadni. Így tehát a kellően pontos $\hat{\mathbb{E}}_{2 \cdot 10^5}(K) = 3360$ becslés adódik. Szóval a várható értékből ítélve azt mondhatjuk, hogy ez még a paraméterek javulását figyelembe véve is egy megdönthetetlen rekord. De sokszor félrevezető lehet a várható érték, hiszen azt a kevésbé valószínű, nagyon nagy értékek megnövelhetik. Az eddigiek alapján könnyen megkaphatjuk, hogy mi a valószínűsége annak, hogy a következő 100 évben megdöntik Usain Bolt rekordját. A számolást elvégezve azt kapjuk, hogy a szóban forgó valószínűség 0.012. Ez az eredmény persze bizonyos fenntartásokkal kezelendő. Ugyanis a nyilvánvaló statisztikai és numerikus pontatlanságok mellett az is előfordulhat, hogy a jövőben megváltozik μ_i alakulásának tendenciája, vagy akár σ és ξ is máshogy viselkedhet a jövőben. Valamint azt sem zárhatjuk ki, hogy a jövőben is lesz valaki, aki nem illeszkedik a mezőny többi részéhez, és más eloszlás szerinti eredményeket produkál.

A következő érdekes kérdés az lehet, hogy mi az emberiség által elérhető legjobb eredmény a 100 méteres síkfutásban. Láttuk, hogy Usain Bolt 9,58 másodperces rekordja szinte megdönthetetlen, szóval egy ahhoz közel eső értékre számítunk. Ha továbbra is feltételezzük a modell jövőbeli helyességét, akkor $\lim_{i \to \infty} \hat{\mu}_i = \hat{\mu}^{(1)}$. Tehát előbb-utóbb a $\hat{\mu}^{(1)}$ jó közelítése lesz a becsült helyparamé-

ternek. Vagyis egy idő után $GEV(\hat{\mu}^{(1)}, \hat{\sigma}, \hat{\xi})$ eloszlással lesz becsülhető az éves legjobb eredmény. Ha ennek az extrémérték-eloszlásnak meghatározzuk a jobb végpontját, akkor megkapjuk az emberiség által elérhető legjobb eredményt. Itt a jobb végpont $\hat{x}_F = \hat{\mu}^{(1)} - \hat{\sigma}/\hat{\xi} = -9,5579$. Tehát körülbelül $9,56$ másodperc a 100 méteres síkfutásban elérhető legjobb idő, azaz Usain Bolt rekordja már csak körülbelül 2 századmásodperccel javítható. Nyilván ez a gondolatmenet is ugyanúgy fenntartásokkal kezelendő, mint a várható érték meghatározásánál elmondottak.

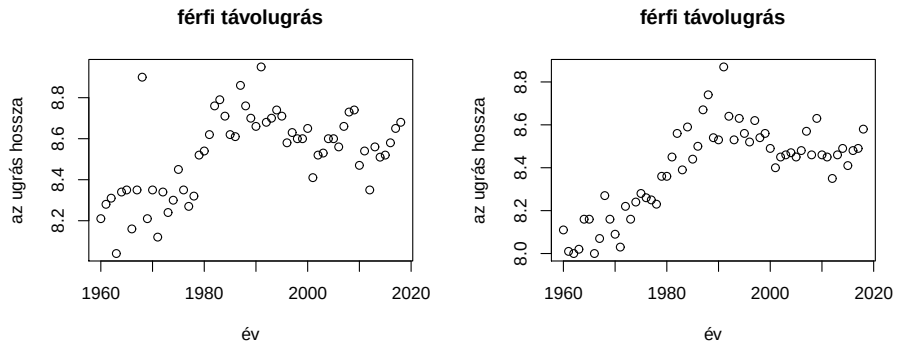
4.1.6. A 400 méteres síkfutás

A 100 méteres férfi síkfutás után tekinthetnénk a 200 méteres síkfutást. Azonban igencsak nagy átfedés tapasztalható ezen két szám között, hiszen sok atléta mindkét számban jeleskedett, például mindkét szám jelenlegi rekordját Usain Bolt tartja. Tehát arra számítunk, hogy nem sok újdonsággal szolgálna a 200 méteres síkfutás modellezése. Ezzel szemben a 400 méteres férfi síkfutás mezőnyében viszonylag kevés olyan atléta található, akik 100 méteren is versenyeztek. Ugyanakkor a 100 méteres síkfutás modellezésénél bemutatott és leellenőrzött modell mégis jól felhasználható a 400 méteres síkfutás elemzése során is. Vagyis az exponenciális modell itt is használható konstans σ és ξ mellett. Az adatok most is ugyanazon a weboldalon, szintén másodpercben vannak megadva.

Most csak röviden ismertetjük a kapott eredményt.
A becsült paraméterek $\hat{\mu}^{(1)} = -43,9008$, $\hat{\mu}^{(2)} = 1,4097$, $\hat{\mu}^{(2)} = 0,0894$, $\hat{\sigma} = 0,2856$, $\hat{\xi} = -0,0920$. Most is adhatunk becslést az emberiség által elérhető legjobb időre. A becsült jobb végpont $\hat{x}_F = \hat{\mu}^{(1)} - \hat{\sigma}/\hat{\xi} = -40,796$. Vagyis az elérhető legjobb idő körülbelül $40,80$ másodperc. A jelenlegi rekord $43,03$ másodperc, melyet Wayde van Niekerk állított be 2016-ban. A 100 méteres síkfutásnál ismertetett módszer segítségével becslést adhatunk arra, hogy várható értékben mikor fog megdőlni ez a rekord. Kellő pontosságig kiszámolva a leírt algoritmussal azt kapjuk, hogy várható értékben $37,0384$ év múlva dől meg a jelenlegi rekord.

4.2. Távolugrás

Érdemes egy olyan atlétikai számmal is foglalkozni, ahol ténylegesen maximumrekord produkálása a cél, és nincs is szükség a (-1)-gyel való szórzásra. A távolugrás egy ilyen atlétikai szám.



(a) $k = 1$ (b) $k = 2$

9. ábra. Az évenkénti k. legjobb eredmények 1960-től 2018-ig

4.2.1. A helyparaméter alakulása

A távolugrás modellezéséhez szintén a [8] weboldalon található adatokat használjuk fel, ugyanúgy a 10 elemű rendezett mintákkal dolgozunk. A rendelkezésre álló adatok méterben vannak megadva. Itt most 1960-tól 2018-ig állnak rendelkezésünkre adatok. Az egyes évek legjobb eredményeit szolgáltató $Y_1, \ldots, Y_{59}$ valószínűségi változókról hasonló dolgok mondhatóak el, mint a síkfutás elemzésénél. Vagyis $Y_i \sim GEV(\mu_i, \sigma_i, \xi_i)$, és feladatunk a (μ_i, σ_i, ξ_i) paraméterek meghatározása. Ismét hivatkozva a [2] könyvben és a [7] cikkben leírtakra, valamint okulva a síkfutás modellezésénél látottakból azt mondhatjuk, hogy a skálaparamétert és az alakparamétert konstansnak érdemes tekinteni.

Vagyis a javuló tendenciát ismét valahogyan μ_i változásával szeretnénk magyarázni. Mielőtt azt tárgyalnánk, hogy a μ_i függvénytől milyen tulajdonságokat várunk el, érdemes ábrázolni az éves legjobb eredményeket. A 9. ábrán láthatjuk a szóban forgó adatokat. Eleinte igencsak szembetűnő javuló tendencia figyelhető meg $k = 1$ és $k = 2$ esetben is. Azonban körülbelül a minta felétől úgy tűnik, hogy romló tendencia lép érvénybe. Ennek a kettősségnek az az oka, hogy a kezdeti fellendülést követően a 90-es évektől kezdve folyamatosan veszít a népszerűségéből a távolugrás. Bővebben erről a jelenségről a [11] linken elérhető cikkben olvashatunk, mely többek között azt nevezi meg a népszerűségcsökkenés indokának, hogy éppen a rövidtávú síkfutás halássza el a potenciális távolugró tehetségeket.

Körülbelül a 30. évben kezdődött el ez a változás, ennek fényében a legkézenfekvőbb jelöltnek most valamilyen $\mu_i = \alpha_i \mathbb{1}_{[1,29]}(i) + \beta_i \mathbb{1}_{[30,59]}(i)$ alakú függvény tűnik a helyparaméter változásának leírására. Ahol α_i és β_i olyan függvények, melyekkel μ_i folytonos. A lehetőségek száma végtelen, és itt most nincs egy

általánosan elfogadott μ_i, mint amilyen a síkfutásnál volt. Így célravezető lehet néhány nem túl sok paramétert használó, de viszonylag rugalmas α_i és β_i konkrét vizsgálata. Nézzünk meg 3 különböző esetet.

Az első $(1a)$ legyen az $\alpha_i = \mu^{(1,1)}i + \mu^{(1,2)}$, $\beta_i = \mu^{(2,1)}i + \mu^{(2,2)}$ eset. Itt ténylegesen csak 3 paraméterünk van $(\mu^{(11)}, \mu^{(12)}, \mu^{(13)})$, hiszen a folytonossági feltételből

$$\mu^{(2,2)} = (\mu^{(1,1)} - \mu^{(2,1)}) \cdot 30 + \mu^{(1,2)}.$$

A második $(2a)$ legyen az $\alpha_i = \mu^{(1,1)}i^2 + \mu^{(1,2)}i + \mu^{(1,3)}$, $\beta_i = \mu^{(2,1)}i^2 + \mu^{(2,2)}i + \mu^{(1,3)}$ eset. Itt a paraméterek számának csökkentésének céljából folytonos differenciálhatóságot is követeljünk meg.
Ekkor 4 paraméterünk van $(\mu^{(11)}, \mu^{(12)}, \mu^{(13)}, \mu^{(21)})$, hiszen

$$\mu^{(2,2)} = 2 \cdot 30(\mu^{(1,1)} - \mu^{(2,1)}) + \mu^{(1,2)},$$

$$\mu^{(2,3)} = 30^2(\mu^{(1,1)} - \mu^{(2,1)}) + 30\Big(\mu^{(1,2)} - \big(2 \cdot 30(\mu^{(1,1)} - \mu^{(2,1)}) + \mu^{(1,2)}\big)\Big) + \mu^{(1,3)}.$$

A harmadik $(3a)$ pedig legyen az $\alpha_i = \mu^{(11)} - \mu^{(12)} \exp(-\mu^{(13)} \cdot i)$, $\beta_i = \mu^{(21)} - \mu^{(22)} \exp(-\mu^{(23)} \cdot i)$ eset. Itt a végtelen sokszor való differenciálhatóságot is elvárhatjuk. Így 4 paraméterünk adódik $(\mu^{(11)}, \mu^{(12)}, \mu^{(13)}, \mu^{(23)})$, hiszen

$$\mu^{(21)} = \mu^{(11)},$$

$$\mu^{(22)} = \mu^{(12)} \exp\big(30(\mu^{(23)} - \mu^{(13)})\big).$$

A síkfutás modellezésénél leírt módon itt is elvégezhető a maximum likelihood becslés. A becslést elvégezve azt tapasztaljuk, hogy a legnagyobb loglikelihood az $(1a)$ modellel adódik, körülbelül $0,2$-vel kisebb érték adódik a $(3a)$ esetben, és jelentősen rosszabb eredmény a $(2a)$ esetben. Az elsőre meglepő lehet, hogy a lineáris modell jobb eredményt produkált a négyzetesnél, hiszen azt hihetnénk, hogy a négyzetes magában foglalja a lineárist. De ez nem így van, ugyanis a négyzetesnél voltak megkötések a függvény simaságára, míg a lineáris esetben csak a folytonosságot követeltük meg. Így az olyan esetekben, ahol éles törés van az eredmények tendenciájában, ott előfordulhat, hogy a lineáris modell lesz a jobb. Fordítva is igaz, ha a lineáris modell jelentősen jobb eredményt produkál, akkor feltételezhetően törés van a helyparaméter változásában. Vagyis helytállónak tűnik az az előzetes sejtésünk, hogy külön kezelendő a vizsgált intervallum két része. Az exponenciális modell 1 paraméterrel többet használ a lineárisnál, és így is egy kicsivel gyengébb eredményt ad, vagyis ez a modell is elvetendő.

Tehát az $(1a)$ modell tűnik a jó választásnak. Nézzük meg a helyparaméter

változását leíró paraméterekre milyen becslést kaptunk. A szóban forgó 3 becsült paraméter $\hat{\mu}^{(11)} = 0,01450$, $\hat{\mu}^{(12)} = 8,1123$, $\hat{\mu}^{(21)} = 0,000934$. Vagyis a törés után μ_i meredeksége rendkívül kicsi, de meglepő módon pozitív. Ennek fényében érdemes lehet a látott 3 modellben a helyparamétert leíró függvényt modosítani $\mu_i = \alpha_i \mathbb{1}_{[1,29]}(i) + \beta \mathbb{1}_{[30,59]}(i)$ alakúra. Ahol α_i olyan, mint eddig, β pedig olyan konstans, mellyel μ_i folytonos. Az így kapott három új modell legyen rendre $(1b)$,$(2b)$ és $(3b)$.Itt már tényleg magában foglalja a négyzetes modell a lineárist. Az így kapott három új modellel számolva a maximum loglikelihood értéke mindhárom esetben szinte ugyanannyi. Konkrétan az $(1a)$ modellnél kapott értéknél körülbelül 1-gyel kisebb értéket kapunk mindhárom esetben.

Így a paraméterek számát figyelembe véve, a három új modell közül $(1b)$ tűnik a legjobbnak. Kérdés, hogy $(1a)$ vagy $(1b)$ a preferálandó modell. Az alakparaméterre kapott becslés a két esetben $\hat{\xi}_a = 0,0101$, míg $\hat{\xi}_b = -0.0050$. Az első fejezetben elmondottak alapján a negatív alakparaméter az elvárt az esetünkben. De igazából a rendkívül kicsi abszolútértékű alakparaméter miatt az $(1b)$ modellből sem adódna észszerű jobb végpont, szóval mindkettő modell csak a kisebb értékekre közelíti jól a valóságot. Mindkét modell tehát elfogadhatónak tűnik, így az egyszerűség kedvéért döntsünk az $(1b)$ modell mellett. Ezen modellel a többi paraméterre pedig $\hat{\mu}^{(1)} = 0,0152$, $\hat{\mu}^{(2)} = 8,1024$, $\hat{\sigma} = 0,1253$ becslés adódik. A síkfutásnál leírt ellenőrzési módszerek itt is elvégezhetőek.

4.2.2. Rekordok a jövőben

Legyen $\hat{\mu} = 30\hat{\mu}^{(1)} + \hat{\mu}^{(2)} = 8,5598$. Tehát az eddigiek alapján $i \geq 30$ esetén $Y_i \sim GEV(\hat{\mu}, \hat{\sigma}, \hat{\xi})$. Feltételezzük, hogy a jövőben is helytálló lesz a megalkotott modell. Ekkor becslést adhatunk az emberiség által elérhető leghosszabb ugrásra. Most azonban a jobb végpont meghatározása a kis abszolút értékű alakparaméter miatt nem tűnik értelmes eljárásnak, hiszen így $\hat{x}_F = \hat{\mu} - \hat{\sigma}/\hat{\xi} = 33,7$ (méter) adódna. Legyen G $GEV(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ eloszlásfüggvény. Ekkor lehető legjobb eredménynek mondjuk egy $G^{-1}(0,99) = 9,1298$ és $G^{-1}(0,999) = 9,4108$ közti érték elfogadható becslésnek tűnik.

Szintén érdemes megnézni, hogy a jelenlegi rekord várható értékben mikor fog megdőlni. A jelenlegi rekord $8,95$ méter, melyet 1991-ben Mike Powell állított be. Ekkor tehát egy évben a rekord megdöntésének valószínűsége $1 - G(8,95) = 0,04246$. Vagyis a rekord várható értékben $\frac{1}{1-G(8,95)} = 23,55$ év múlva dől meg.

Összefoglalás

A szakdolgozat első fejezete során betekintést nyerhettünk az extrémérték-elmélet alapjaiba. Meggondoltuk, hogy egy adott sportág legjobb eredményeinek vizsgálatakor hogyan érdemes alkalmazni a Fisher és Tippet nevéhez fűződő főtételt.

A második fejezet során az első fejezetben látott főtételnek egy fontos általánosításával ismerkedhettünk meg. Ezen általánosítás a későbbiekben egy rendkívül jól használható eszköznek bizonyult a rendezett minták elemzése során.

A harmadik fejezetben láttuk, hogy független, azonos eloszlású, abszolút folytonos valószínűségi változókból rendkívül ritkán adódik rekord. A fejezetben kimondott állítások segítségével láttuk, hogy a 100 méteres síkfutás évenkénti legjobb idejeit megadó valószínűségi változók nem tekinthetőek független, azonos eloszlásúnak.

A negyedik fejezetben pedig a 100 méteres síkfutás, a 400 méteres síkfutás és a távolugrás legjobb eredményei kerültek modellezésre. Előrejelzést adtunk az éppen aktuális rekordok megdöntésével és a valaha elérhető legjobb eredményekkel kapcsolatban. Kiemelendő a 100 méteres síkfutás modellezésénél kapott konklúzió, miszerint Usain Bolt rekordja a jövőben könnyen megdönthetetlennek bizonyulhat.

A téma további vizsgálata is lehetséges. Egyrészt elméleti szempontból is kiegészíthető a dolgozat, hiszen az első fejezet során kimondásra kerülő tételek bizonyítása is érdekes lehet. Valamint a többváltozós extrémérték-elmélet is egy elméleti szempontból fontos témakör, melynek ismertetésére ezen dolgozat során nem került sor.

A modellezés mélységét illetően is adódik lehetőség a fejlesztésre. Például együtt tekinthetjük a férfi és női eredményeket, és vizsgálhatjuk a két nem által produkált eredmények függetlenségének hipotézisét. Ha azt kapjuk, hogy az egyes évek legjobb női és férfi eredményei összefüggenek, akkor a többváltozós extrémérték-elmélet eszközeihez is folyamodhatunk. Erről a gondolatmenetről bővebben a [10] cikkben olvashatunk. Hasonlóan járhatunk el több atlétikai

szám együttes vizsgálatával. Itt az összefüggés még egyértelműbb, mint a nemek vizsgálatánál, hiszen vannak atléták, akik több atlétikai számban is kimagasló eredményt értek el.

Irodalomjegyzék

[1] Paul Embrechts, Claudia Klüppelberg, Thomas Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag Berlin Heidelberg, 1997.

[2] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag London, 2001.

[3] R.-D. Reiss. *Approximate Distributions of Order Statistics*. Springer-Verlag New York, 1989.

[4] Sidney I. Resnick. *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag New York, 1987.

[5] M. R. Leadbetter, Georg Lindgren, Holger Rootzén. *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag New York, 1983.

[6] Sneh Gulati, W. J. Padgett. *Parametric and Nonparametric Inference from Record-Breaking Data*. Springer-Verlag New York, 2003.

[7] M. B. Adam, J. A. Tawn. *Modelling Record Times in Sport with Extreme Value Methods*. Malaysian Journal of Mathematical Science, 2016.

[8] `http://www.iaaf.org`

[9] A. L. M. Dekkers, J. H. J. Einmahl, L. De Haan. *A Moment Estimator for the Index of an Extreme-Value Distribution* . The Annals of Statistics, 1989.

[10] M. B. Adam, J. A. Tawn.. *Bivariate Extreme Analysis of Olympic Swimming Data* . Journal of Statistical Theory and Practice, 2012.

[11] `https://slate.com/culture/2012/08/long-jump-olympics-why-do-the-best-long-jumpers-in-the-world-seem-to-be-jumping-shorter-distances.html`