Approximációs tételek a kupongyűjtő problémában

Doktori (Ph.D.) értekezés tézisei

Pósfai Anna

Témavezetők:
Dr. Csörgő Sándor
egyetemi tanár
és
Dr. Andrew D. Barbour
egyetemi tanár

Matematika- és Számítástudományok Doktori Iskola
Szegedi Tudományegyetem
Bolyai Intézet

2010

1. Bevezetés

1.1. A kupongyűjtő probléma

A kupongyűjtő probléma a valószínűségszámítás egyik klasszikus problémája. A probléma legegyszerűbb és minden bizonnyal eredeti megfogalmazása a következő: Tegyük fel, hogy n darab kuponból véletlenszerűen visszatevéssel húzunk. Mennyi annak a valószínűsége, hogy több, mint t húzásra van szükségünk ahhoz, hogy megszerezzük mind az n kupont. A probléma egyik legkorábbi tárgyalása Pólya nevéhez fűződik [14]. Feller munkájában hétszer említik [9]. A problémának számos változata és általánosítása létezik. Szoros kapcsolatban áll az urna problémákkal, illetve különböző véletlen jelenségek várakozási idejének vizsgálatával (pl. [12], [11], [1]).

A dolgozatban a probléma következő változatával foglalkozunk: adva van $n \geq 2$ különböző kupon, melyekből egy gyűjtő véletlenszerű visszatevéses mintát vesz úgy, hogy minden egyes alkalommal az n kupon bármelyikét azonos, tehát $1/n$ valószínűséggel szerzi meg. Valamely rögzített $m \in \{0, 1, \ldots, n-1\}$ esetén a mintavételt addig folytatja, amíg először pontosan $n - m$ különböző kupont nem gyűjtött. Jelölje $W_{n,m}$ az ehhez szükséges ismétlések számát. Tehát a $W_{n,m}$ véletlen változó, amelyet a kupongyűjtő várakozási idejének nevezünk, az $n-m, n-m+1, n-m+2, \ldots$ értékeket veheti fel, és megadja, hogy a gyűjtőnek hányszor kell húznia ahhoz, hogy m darab kupon kivételével minden kupon a birtokában legyen. Speciálisan, $W_{n,0}$ a teljes gyűjtemény megszerzéséhez szükséges várakozási idő.

A dolgozatban a várakozási idő várható értékét $\mu_n = \mu_n(m) := \mathbf{E}(W_{n,m})$, szórásnégyzetét pedig $\sigma_n^2 = \sigma_n^2(m) := \mathbf{Var}(W_{n,m})$ jelöli.

1.2. Határeloszlás tételek a kupongyűjtő problémában

Különböző határeloszlás tételeket bizonyítottak $W_{n,m}$ aszimptotikus eloszlására attól függően, hogy az $m = m(n)$ sorozat hogyan viselkedik, amint $n \to \infty$. A továbbiakban minden konvergencia és aszimptotikus reláció $n \to \infty$ mellett értendő.

Az első eredmény Erdős és Rényi [8] nevéhez fűződik, akik egy eltolt Gumbel-eloszlást kaptak határeloszlásként a teljes gyűjtemény esetére, amikor m minden $n \in \mathbb{N}$ esetén 0. Ezt az eredményt általánosította Baum és Billingsley [6], akik minden $m = m(n)$ sorozat típust vizsgáltak. Négy különböző határeloszlást határoztak meg:

1. Elfajuló eloszlás

$$\text{Ha } \frac{n-m}{\sqrt{n}} \to 0, \text{ akkor } W_{n,m} - (n-m) \xrightarrow{\mathcal{D}} 0, \tag{1}$$

azaz a limesz valószínűségi mérték a 0-ra koncentrált.

2. **Poisson eloszlás**

$$\text{Ha } \frac{n-m}{\sqrt{n}} \to \sqrt{2\lambda}, \text{ akkor } W_{n,m} - (n-m) \xrightarrow{\mathcal{D}} \mathrm{Po}(\lambda), \tag{2}$$

ahol $\mathrm{Po}(\lambda)$ a λ paraméteres Poisson eloszlás: $\mathrm{Po}(\lambda)\{k\} = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}$, $k = 0, 1, 2, \ldots$.

3. **Normális eloszlás**

$$\text{Ha } \frac{n-m}{\sqrt{n}} \to \infty \text{ és } m \to \infty, \text{ akkor } \frac{W_{n,m} - \mu_n}{\sigma_n} \xrightarrow{\mathcal{D}} \mathrm{N}(0,1), \tag{3}$$

ahol $\mathrm{N}(0,1)$ a standard normális eloszlás, amelynek Lebesgue mértékre vonatkozó sűrűségfüggvénye $\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$, $x \in \mathbb{R}$.

4. **Gumbel-típusú eloszlás**

$$\text{Ha } m \equiv \text{konstans}, \text{ akkor } \frac{W_{n,m} - \mu_n}{n} \xrightarrow{\mathcal{D}} \mathrm{Gumbel}(m), \tag{4}$$

ahol $\mathrm{Gumbel}(m)$ a későbbi (5) formula segítségével definiált valószínűségi mérték, melyet m paraméteres Gumbel-típusú eloszlásnak nevezünk.

1.3. Célkitűzések

A dolgozatban finomítjuk a fenti, Baum és Billingsley nevéhez fűződő határeloszlás tételeket. Célunk a megfelelően centralizált és normalizált várakozási idő eloszlásának jól ismert mértékekkel történő minél pontosabb közelítése, és sok esetben a kapcsolódó eloszlásfüggvények és valószínűségi pontfüggvények aszimptotikus sorfejtése. A közelítő mértékeket öt különböző mértékcsaládból választjuk. Ezek közül három – a Poisson eloszlások, a normális eloszlások és a Gumbel-típusú eloszlások – olyan mértékcsaládok, melyeknek tagjai határeloszlásaként szerepelnek Baum és Billingsley határeloszlás tételeiben.

A negyedik approximáló mértékcsalád az összetett Poisson-eloszlásoknak bizonyos $\{\pi_{\mu,a} : \mu > 0, a > 0\}$ osztálya. Tetszőleges $\mu > 0$ és $a > 0$ esetén $\pi_{\mu,a}$ a $Z_1 + 2Z_2$ véletlen változó eloszlását jelöli, ahol Z_1 és Z_2 valamely közös valószínűségi mezőn definiált független véletlen változók, $Z_1 \sim \mathrm{Po}(\mu)$ és $Z_2 \sim \mathrm{Po}(a/2)$. Valószínűségi generátorfüggvényének segítségével könnyen ellenőrizhető, hogy $\pi_{\mu,a}$ valóban összetett Poisson eloszlás, azaz eloszlásban egyenlő egy $\sum_{k=1}^{N} X_k$ alakú véletlen változóval, ahol $N, X_1, X_2, \ldots$ közös valószínűségi mezőn értelmezett független véletlen változók, melyek közül N Poisson eloszlású, $X_1, X_2, \ldots$ pedig azonos eloszlásúak.

Az ötödik közelítő mértékcsalád a Poisson-Charlier előjeles mértékek osztálya. Tetszőleges pozitív valós λ, $\widetilde{a}^{(1)}, \ldots, \widetilde{a}^{(S)}$ és $S \in \mathbb{N}$ esetén a $\nu = \nu(\lambda, \widetilde{a}^{(1)}, \ldots, \widetilde{a}^{(S)})$ Poisson-Charlier előjeles mérték az az előjeles mérték, amely a nemnegatív egészekre van koncentrálva, és

$$\nu\{j\} = \mathrm{Po}\{j\}(\lambda) \left(\sum_{r=1}^{S} (-1)^r \widetilde{a}^{(r)} C_r(j,\lambda) \right), \quad j \in \mathbb{N},$$

ahol $C_r(j,\lambda)$ az r-edik Charlier polinom ([7] 170. o.).

2. Módszerek valószínűségi mértékek közelségének mérésére

Legyenek μ és ν az $(\mathbb{R},\mathcal{B})$ mérhetőségi téren értelmezett valószínűségi mértékek, ahol $\mathcal{B}$ a valós egyenes Borel-halmazainak σ-algebráját jelöli. A dolgozatban a két eloszlás közelségének mérésére a

$$d_{\mathrm{K}}(\mu,\nu)=\sup_{x\in\mathbb{R}}\big|\mu((-\infty,x])-\nu((-\infty,x])\big|$$

Kolmogorov távolságot és a

$$d_{\mathrm{TV}}(\mu,\nu)=\sup_{B\in\mathcal{B}}\big|\mu(B)-\nu(B)\big|$$

teljes variációs távolságot használjuk.

A dolgozat bizonyításaiban használt fő eszközök között szerepelnek különböző karakterisztikus függvényeket alkalmazó technikák, a Stein-módszer, csatolásos módszerek és bizonyos elemi kombinatorikai megfontolások.

3. Gumbel-típusú approximáció

Ebben a fejezetben az

$$F_{n,m}(x):=\mathbf{P}\left(\frac{1}{n}W_{n,m}-\sum_{k=m+1}^{n}\frac{1}{k}\le x\right),\quad x\in\mathbb{R},$$

eloszlásfüggvény aszimptotikus viselkedését vizsgáljuk abban az esetben, amikor m minden n-re rögzített konstans és $n\to\infty$.

A (4) pontbeli gyenge konvergencia felírható

$$\lim_{n\to\infty}F_{n,m}(x)=F_m(x):=\frac{1}{m!}\int_{-\infty}^{x}\mathrm{e}^{-(m+1)(y+C_m)}\,\mathrm{e}^{-\mathrm{e}^{-(y+C_m)}}\mathrm{d}y,\quad x\in\mathbb{R},\tag{5}$$

alakban, ahol $C_m:=\gamma-\sum_{k=1}^{m}\frac{1}{k}$ és $\gamma=\lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{1}{k}-\log n\right)=0,577\ldots$.

A disszertációban minden m esetén olyan $F_m+G_{n,m}$ egytagú aszimptotikus sorfejtést adunk, amely egyenletesen $1/n$ rendben közelíti az $F_{n,m}$ eloszlásfüggvényt, továbbá az explicit módon megadott $G_{n,m}$ függvények sorozata egyenletesen $(\log n)/n$ rendű. Ezáltal speciálisan azt is bizonyítjuk, hogy (5)-ben a konvergenciasebesség rendje $(\log n)/n$.

Minden $n\ge m+2$ esetén bevezetjük a

$$G_{n,m}(x)=-\frac{1}{2n}\sum_{k=m+1}^{n-1}\frac{1}{k}\int_{-\infty}^{x}[f_m''\star h_k](u)\,\mathrm{d}u,\quad x\in\mathbb{R},$$

függvényeket, ahol $f_m(x) := F'_m(x)$ a határeloszlás sűrűségfüggvénye, $h_k(x) = \mathrm{e}^{-kx}$, $x > 0$, az $1/k$ várható értékű exponenciális eloszlás sűrűségfüggvénye, és $\star$ a konvolúciót jelöli. Fő eredményünk a

3.1.1. Tétel. *Bármely rögzített $m \in \{0, 1, 2, \ldots\}$ esetén*

$$\sup_{x\in\mathbb{R}} \bigl|F_{n,m}(x) - [F_m(x) + G_{n,m}(x)]\bigr| = O\Bigl(\frac{1}{n}\Bigr), \tag{6}$$

továbbá a $G_{n,m}$ függvényekhez létezik olyan m-től függő $K_m > 0$ konstans, $x_m \in \mathbb{R}$ pont, $c_m(\cdot)$ pozitív függvény és $n_m(\cdot) \in \mathbb{N}$ küszöbfüggvény, hogy

$$\sup_{x\in\mathbb{R}} \bigl|G_{n,m}(x)\bigr| \le K_m \frac{\log n}{n}, \quad n \ge m+2,$$

ugyanakkor

$$\bigl|G_{n,m}(x)\bigr| \ge c_m(x) \frac{\log n}{n}, \quad x \in (-\infty, x_m),$$

valahányszor $n \ge n_m(x)$.

A dolgozatban azt is belátjuk, hogy (6)-ban a hibarend éles, és hogy semmilyen a (6) pontban megadott sorfejtésnél hosszabb aszimptotikus sorfejtés esetén nem kaphatunk az $1/n$ rendnél kisebb hibarendet.

A fejezet eredményei publikálásra kerültek [19]-ben.

4. Normális approximáció

A dolgozat ezen fejezetében felső becslést adunk a kupongyűjtő várakozási idejének normális eloszlással való közelítésének hibájára. Bevezetjük az

$$F_{n,m}(x) := \mathbf{P}\biggl(\frac{W_{n,m} - \mu_n}{\sigma_n} \le x\biggr), \quad x \in \mathbb{R},$$

eloszlásfüggvényt, és igazoljuk az alábbi tételt.

4.0.1. Tétel. *Minden $n \ge 3$ és $1 \le m \le n-2$ esetén*

$$\sup_{x\in\mathbb{R}} \bigl|F_{n,m}(x) - \Phi(x)\bigr| \le C \frac{n}{m} \frac{1}{\sigma_n},$$

ahol Φ a standard normális eloszlásfüggvény és $C = 9.257$.

Ellenőrizhető, hogy a 4.0.1. Tétel által adott felső becslés pontosan akkor tart nullához, ha m végtelenbe megy, de elég lassan ahhoz, hogy a $(n-m)/\sqrt{n}$ sorozat is végtelenbe tartson, ami éppen a (3) határeloszlást adja.

A fejezet eredményei publikálásra kerültek [18]-ban.

5. Poisson approximáció

5.1. Poisson approximáció egy általános Poisson határeloszlás tételben

Az 5. fejezet első alfejezetében általános független nemnegatív egészértékű véletlen változók összegeinek eloszlását közelítjük Poisson eloszlással. Az alábbi Gnedenko és Kolmogorov [10] (132. o.) nevéhez fűződő Poisson határeloszlás tételt finomítjuk.

5.1.1. Tétel. (Gnedenko, Kolmogorov) *Legyen $\{Y_{n1}, Y_{n2}, \ldots, Y_{nr_n}\}_{n\in\mathbb{N}}$ soronként független, nemnegatív egészértékű véletlen változókból álló szériasorozat, melyre*

$$\min_{1\le k\le r_n} \mathbf{P}(Y_{nk}=0)\to 1, \quad n\to\infty,$$
$$\sum_{k=1}^{r_n} \mathbf{P}(Y_{nk}\ge 1)\to\lambda, \quad \lambda>0 \textit{ konstans}, \quad n\to\infty$$
$$\sum_{k=1}^{r_n} \mathbf{P}(Y_{nk}\ge 2)\to 0, \quad n\to\infty.$$

Ekkor

$$Y_n := \sum_{k=1}^{r_n} Y_{nk} \xrightarrow{\mathcal{D}} \mathrm{Po}(\lambda), \quad \textit{amint } n\to\infty.$$

Tetszőleges soronként független, nemnegatív egészértékű véletlen változókból álló szériasorozat tekintünk, és minden n esetén az n-edik sorösszeg eloszlását olyan Poisson eloszlással közelítjük, melynek λ_n várható értéke csak az adott sorban szereplő változók eloszlásától függ, de nem követeljük meg momentumok létezését:

$$\lambda_n = \sum_{k=1}^{r_n} \mathbf{P}(Y_{nk}\ge 1).$$

A közelítés hibájára alsó és felső korlátot is adunk, melyek rendje konstans szorzótól eltekintve megegyezik, feltéve, hogy a λ_n paraméterek korlátosak. Ezáltal jobb közelítését adjuk az Y_n változóknak, mint amit a kézenfekvő, határeloszlással történő approximáció jelent.

5.1.2. Tétel. (Felső korlát.) *Ha $\{Y_{n1}, Y_{n2}, \ldots, Y_{nr_n}\}_{n\in\mathbb{N}}$ soronként független, nemnegatív egészértékű véletlen változókból álló szériasorozat, akkor*

$$d_{\mathrm{TV}}(\mathcal{D}(Y_n), \mathrm{Po}(\lambda_n)) \le \sum_{k=1}^{r_n} \left[\mathbf{P}(Y_{nk}\ge 2) + \mathbf{P}(Y_{nk}\ge 1)^2\right].$$

5.1.3. Tétel. (Alsó korlát.) *Ha* $\{Y_{n1}, Y_{n2}, \ldots, Y_{nr_n}\}_{n\in\mathbb{N}}$ *olyan soronként független, nemnegatív egészértékű véletlen változókból álló szériasorozat, hogy* $\min_{1\le k\le r_n} \mathbf{P}(Y_{nk} = 0) \ge \frac{3}{4}$ *minden* $n \in \mathbb{N}$ *esetén, akkor*

$$d_{\mathrm{TV}}(\mathcal{D}(Y_n), \mathrm{Po}(\lambda_n)) \ge \frac{1}{10}\left(\prod_{k=1}^{r_n} \mathbf{P}(Y_{nk} = 0)\right) \sum_{k=1}^{r_n} \left[\mathbf{P}(Y_{nk} \ge 2) + \mathbf{P}(Y_{nk} \ge 1)^2\right].$$

Az 5.1.2. és 5.1.3. Tételekből következik, hogy az 5.1.1. Tételben szereplő véletlen változók fenti Poisson approximációja esetén a hiba rendje

$$\sum_{k=1}^{r_n} \left[\mathbf{P}(Y_{nk} \notin \{0,1\}) + \mathbf{P}(Y_{nk} \ge 1)^2\right].$$

Barbour és Hall a Stein-módszer segítségével hasonló eredményeket igazoltak [3]-ban: független nemnegatív egészértékű véletlen változók $\sum_{j=1}^n Y_j$ összegeit olyan Poisson eloszlású véletlen változóval közelítették, melynek várható értéke $\sum_{j=1}^n \mathbf{P}(Y_j = 1)$ vagy $\sum_{j=1}^n \mathbf{E}(Y_j)$. (Megjegyezzük, hogy az általunk alkalmazott approximációban a közelítő Poisson eloszlás paramétere e két érték között van.) Az általuk megadott korlátok más alakúak és szerepelnek bennük az Y_j változók második momentumai. Továbbá alsó korlátaik sokkal bonyolultabb kifejezések, melyek nem adnak használható információt abban az esetben, amikor a kupongyűjtő várakozási idejére alkalmazzuk őket.

5.1.1. Következmény. *Az 5.1.1. Tételben szereplő határeloszlásban a konvergencia sebessége*

$$d_{\mathrm{TV}}(\mathcal{D}(Y_n), \mathrm{Po}(\lambda)) \le \sum_{k=1}^{r_n} \left[\mathbf{P}(Y_{nk} \ge 2) + \mathbf{P}(Y_{nk} \ge 1)^2\right] + \left|\sum_{k=1}^{r_n} \mathbf{P}(Y_{nk} \ge 1) - \lambda\right|.$$

5.2. Kupongyűjtés közel Poisson eloszlású várakozási idő esetén – az általános eredmények alkalmazása

Ezen alfejezetben először megvizsgáljuk, hogy a kupongyűjtő probléma hogyan illeszkedik az előző alfejezetbeli felálláshoz. Belátható, hogy a $\widetilde{W}_{n,m} := W_{n,m} - (n-m)$ eltolt várakozási időre teljesül a

$$\widetilde{W}_{n,m} \stackrel{\mathcal{D}}{=} \sum_{i=m+1}^{n} \widetilde{X}_{n,i} \tag{7}$$

eloszlásbeli egyenlőség, ahol az $\widetilde{X}_{n,i}$ véletlen változók függetlenek, és $\widetilde{X}_{n,i} + 1$ geometriai eloszlású i/n siker-valószínűséggel, $i \in \{m+1, \ldots, n\}$, $n \in \mathbb{N}$. Ellenőrizhető, hogy a $\{\widetilde{X}_{n,m+1}, \ldots, \widetilde{X}_{n,n}\}_{n\in\mathbb{N}}$ szériasorozat kielégíti az 5.1.1. Tétel feltételeit. Látjuk tehát, hogy a (2) határeloszlás tétel speciális esete a Gnedeno–Kolmogorov tételnek. Ha az előző alfejezet eredményeit alkalmazzuk $\widetilde{W}_{n,m}$-re, a következőket kapjuk.

5.2.1. Következmény *Ha $\{m = m(n)\}_{n\in\mathbb{N}}$ olyan egészekből álló sorozat, amely teljesíti a (2) Poisson határeloszlás tétel feltételeit, akkor a kupongyűjtő $\widetilde{W}_{n,m}$ eltolt várakozási idejének $\lambda_n = \sum_{i=m+1}^{n}\left(1-\frac{i}{n}\right)$ paraméteres Poisson eloszlású N_{λ_n} véletlen változóval történő közelítése esetén a hiba rendje $\sum_{i=m+1}^{n}\left(1-\frac{i}{n}\right)^2$. Sőt, ha $\min_{m+1\le i\le n}\frac{i}{n}\ge\frac{3}{4}$ minden n-re, akkor*

$$\frac{1}{5}\left(\prod_{i=m+1}^{n}\frac{i}{n}\right)\sum_{i=m+1}^{n}\left(1-\frac{i}{n}\right)^2 \le d_{\mathrm{TV}}(\mathcal{D}(\widetilde{W}_{n,m}),\mathcal{D}(N_{\lambda_n})) \le 2\sum_{i=m+1}^{n}\left(1-\frac{i}{n}\right)^2.$$

5.2.2. Következmény *A (2) Poisson határeloszlás tételben a konvergenciasebesség rendjére a következő felső korlát adható:*

$$d_{\mathrm{TV}}(\mathcal{D}(\widetilde{W}_{n,m}),\mathcal{D}(N_{\lambda})) \le 2\sum_{i=m+1}^{n}\left(1-\frac{i}{n}\right)^2 + \left|\sum_{i=m+1}^{n}\left(1-\frac{i}{n}\right)-\lambda\right|.$$

A fejezet első két alfejezetének eredményei publikálásra kerültek [17]-ben.

5.3. Kupongyűjtés közel Poisson eloszlású várakozási idő esetén – kombinatorikai megközelítés

A dolgozat ezen alfejezetében a kupongyűjtő probléma kombinatorikai struktúrájára építünk. Egy kombinatorikai megfontolásokra támaszkodó módszer segítségével erősebb eredményt tudunk igazolni, mint az 5.2.1. Következményben: a $\mathbf{P}(\widetilde{W}_{n,m}=k)$, $k=0,1,\ldots$, valószínűségek megfelelő Poisson valószínűségekkel történő közelítését pontosítjuk az első korrekciós tag meghatározása révén. Az eredményt a 5.3.1. Tétel tartalmazza. Megjegyezzük, hogy a bizonyításban alkalmazott módszer elméletben az aszimptotikus sorfejtés magasabb rendű tagjainak meghatározására is alkalmas.

Legyen

$$\lambda_{n,j} := \sum_{i=m+1}^{n}\left(1-\frac{i}{n}\right)^j, \quad j=1,2,\ldots. \tag{8}$$

Belátható, hogy ha $\{m = m(n)\}_{n\in\mathbb{N}}$ olyan egészekből álló sorozat, amely teljesíti a (2) Poisson határeloszlás tétel feltételeit, akkor $\lambda_n \to \lambda$ és $\lambda_{n,j}\to 0$, $j=2,3,\ldots$, továbbá $\lambda_{n,2} = \frac{(2\lambda_n)^{3/2}}{3\sqrt{n}} + O\left(\frac{1}{n}\right)$.

5.3.1. Tétel. *Ha $\{m = m(n)\}_{n\in\mathbb{N}}$ olyan egészekből álló sorozat, amely teljesíti a (2) Poisson határeloszlás tétel feltételeit, és λ_n valamint $\lambda_{n,2}$ a (8) formulával vannak definiálva,* *akkor*

$$\mathbf{P}(\widetilde{W}_{n,m} = 0) = \mathrm{e}^{-\lambda_n} - \mathrm{e}^{-\lambda_n}\frac{\lambda_{n,2}}{2} + O\left(\frac{1}{n}\right),$$
$$\mathbf{P}(\widetilde{W}_{n,m} = 1) = \mathrm{e}^{-\lambda_n}\lambda_n - \mathrm{e}^{-\lambda_n}\lambda_n\frac{\lambda_{n,2}}{2} + O\left(\frac{1}{n}\right),$$
$$\mathbf{P}(\widetilde{W}_{n,m} = k) = \mathrm{e}^{-\lambda_n}\frac{\lambda_n^k}{k!} + \mathrm{e}^{-\lambda_n}\left(\frac{\lambda_n^{k-2}}{(k-2)!} - \frac{\lambda_n^k}{k!}\right)\frac{\lambda_{n,2}}{2} + O\left(\frac{1}{n}\right), \quad k \geq 2.$$

Ezen alfejezet eredményei szerepelnek [17]-ben, a részleteket [16] tartalmazza.

5.4. Poisson approximáció – a várható érték illesztése

Az 5. fejezet utolsó alfejezetében a kupongyűjtő $\widetilde{W}_{n,m} = W_{n,m} - (n-m)$ eltolt várakozási idejét egy újabb Poisson eloszlású véletlen változóval közelítjük, méghozzá olyannal, amelynek várható értéke megegyezik $\widetilde{W}_{n,m}$ várható értékével. Kiszámolható, hogy azon n és m paraméterértékek esetén, melyekre érvényes a (2) Poisson határeloszlás tétel, az approximáció hibája a lenti 5.4.1. Tételben $1/n$, ami kisebb, mint az ugyanezen esetben az 5.2.1. Következményben vagy az 5.3.1. Tételben bizonyított $1/\sqrt{n}$-es hibarend. Az 5.4.1. Tétel bizonyítása a Stein-módszeren alapszik, és kihasználja azt a tényt, hogy az összehasonlított eloszlások várható értékei egyenlőek. Megjegyezzük, hogy az 5.1.2. Tétel bizonyításában alkalmazott érvelés ebben az esetben nem működne.

5.4.1. Tétel *Ha* $\lambda_n' = \mathbf{E}(\widetilde{W}_{n,m}) = \sum_{i=m+1}^{n}\left(\frac{n}{i} - 1\right)$*, akkor*

$$d_{\mathrm{TV}}(\mathcal{D}(\widetilde{W}_{n,m}), \mathrm{Po}(\lambda_n')) \leq 8\left(1 \wedge \sqrt{\frac{2}{\mathrm{e}\lambda_n'}}\right)\sum_{i=m+1}^{n}\left(\frac{n-i}{i}\right)^3.$$

6. Összetett Poisson approximáció

6.1. A Mineka csatolási egyenlőtlenség egy általánosítása

Igen nagy irodalma van a független egészértékű véletlen változók összegeinek összetett Poisson approximációjának. A Stein-módszer segítségével az [5] és [2] dolgozatok ilyen típusú approximációk teljes variációs távolságban mért hibáira adnak felső korlátokat. Ezek a korlátok az $X_1, X_2, \ldots, X_n$ összeadandók első három momentumának és a $d_{\mathrm{TV}}\left(\mathcal{D}\left(W_n\right), \mathcal{D}\left(W_n + 1\right)\right)$ kifejezésnek függvényei, ahol $W_n = \sum_{j=1}^{n} X_j$.

A $d_{\mathrm{TV}}\left(\mathcal{D}\left(W_n\right), \mathcal{D}\left(W_n + 1\right)\right)$ kifejezést általában a Mineka-csatolás [13] segítségével lehet becsülni, ami tipikusan $1/\sqrt{n}$ rendű eredményt ad. Ha az X_j véletlen változók nagyjából azonos szórásúak, akkor ez az eredmény közel van a centrális határeloszlás tétel esetén elvárt $O(1/\sqrt{\mathbf{Var}W_n})$ hibarendhez. Azonban ha az X_j véletlen változók eloszlásai egyre laposabbak, amint j nő, akkor $\mathbf{Var}W_n$ nőhet gyorsabban, mint n, és ekkor $1/\sqrt{n}$ jóval nagyobb lesz, mint az elvárható $1/\sqrt{\mathbf{Var}W_n}$-es rend. Pontosan ez a helyzet, ha W_n-nek a kupongyűjtő várakozási idejét választjuk.

6.1.1. Lemma. *Legyenek $U_1, U_2, \ldots, U_r$, $r \geq 2$, független azonos eloszlású véletlen változók, melyek valamely $l \geq 1$ esetén egyenletes eloszlásúak az $\{1, 2, \ldots, 2l-1, 2l\}$ véges halmazon. Ha $V_r = \sum_{j=1}^{r} U_j$, akkor*

$$d_{\mathrm{TV}}(\mathcal{D}(V_r), \mathcal{D}(V_r + 1)) \leq \frac{1}{l\sqrt{r}}.$$

A 6.1.1. Lemma megjavítja a Mineka-csatolás által ugyanerre a kifejezésre adott $1/(\sqrt{2r})$ korlátot. A 6.1.1. Állítás segítségével megmutatjuk, hogy a független diszkrét egyenletes eloszlású véletlen változókra vonatkozó lemma eredménye hogyan alkalmazható tetszőleges független egészértékű véletlen változók összegeire vonatkozó hasonló eredmények bizonyítására. Az alapötelet az, hogy az egyenletes eloszlású változókat ágyazzuk be azokba, amelyekre az eredményt igazolni szeretnénk.

6.1.1. Állítás. *Ha $X_1, X_2, \ldots, X_n$, $n \geq 2$, független egészértékű véletlen változók és $W = \sum_{j=1}^{n} X_n$, akkor*

$$d_{\mathrm{TV}}(\mathcal{D}(W), \mathcal{D}(W+1)) \leq \frac{4}{l\sqrt{nlp}} + \frac{8d_n}{nlp},$$

ahol $l \in \{2, 4, 6, \ldots\}$ valamint $p \leq \min\{\mathbf{P}(X_j = k) : k = 1, \ldots, l, j = 1, \ldots, n\}$ tetszőlegesek és $d_n = d_{\mathrm{TV}}(\mathcal{D}(X_n), \mathcal{D}(X_n + 1))$.

A dolgozatban megállapítjuk, hogy az állításban az általánosság megszorítása nélkül feltehettük, hogy az l-intervallumok 1-nél kezdődnek, és hogy a (p, l) paraméterek megválasztása nagymértékben függ az adott problémától. Azt is megjegyezzük, hogy az állítás által adott felső korlátban szereplő konstans javítható a bizonyításban alkalmazott módszer finomításával: az X_j véletlen változókba nemcsak egy, hanem több diszkrét egyenletes eloszlású változót is be lehet ágyazni, ha a valós egyenest felbontjuk l-hosszú intervallumokra, és ezen $(\{(m-1)l, \ldots, ml\})_{m \in \mathbb{Z}}$ blokkok mindegyikéhez definiálunk egy egyenletes eloszlású változót.

6.2. Összetett Poisson approximáció a centrális és Poisson határeloszlás tételek tartományában

Ebben az alfejezetben a kupongyűjtő megfelelően centralizált várakozási idejének eloszlását a bevezetésben definiált $\pi_{\mu,a}$ összetett Poisson eloszlással közelítjük. A (7) eloszlásbeli egyenlőségre alapozva, független egészértékű véletlen változók összegeire vonatkozó általános összetett Poisson approximációs eredményeket és az új csatolásunkat alkalmazzuk. Belátjuk, hogy a $W_{n,m}$ várakozási idő jól közelíthető összetett Poisson eloszlással abban az esetben, amikor az n és m paraméterek teljesítik a (3) centrális vagy a (2) Poisson határeloszlás tétel feltételeit.

6.2.1. Tétel. *Bármely rögzített $n \geq 2$ és $2 \leq m \leq n-4$ esetén ha*

$$\begin{aligned} \mu &= \sigma_n^2 - 2\langle \sigma_n^2 - \mu_n \rangle, \\ a &= \langle \sigma_n^2 - \mu_n \rangle \quad \textit{és} \\ c &= \lfloor \sigma_n^2 - \mu_n \rfloor, \end{aligned}$$

ahol $\langle x \rangle$, illetve $\lfloor x \rfloor$ az x valós szám tört-, illetve egészrészét jelölik, akkor létezik olyan pozitív C konstans, hogy

$$d_{\mathrm{TV}}\Big(\mathcal{D}\left(W_{n,m}+c\right), \pi_{\mu,a}\Big) \leq \frac{C}{\sigma_n}\left(\frac{\lfloor \sigma_n^2 - \mu_n - (n-m) \rfloor}{\sigma_n^2} + \frac{(n-m)^2}{nm}\right).$$

A 6.2.1. Tétel eredményeit a μ_n, σ_n és $a_{n,2} := \sigma_n^2 - \mu_n - (n-m)$ mennyiségekre vonatkozó aszimptotikus formulák segítségével egyszerűbb alakban is kifejezhetjük:

$$d_{\mathrm{TV}}\Big(\mathcal{D}\Big(W_{n,m}+c\Big), \pi_{\mu,a}\Big) = \begin{cases} O\Big(\frac{1}{\sqrt{m}}\Big), & \text{ha } \frac{m}{n} \to 0; \\ O\Big(\frac{1}{\sqrt{n}}\Big), & \text{ha } \frac{m}{n} \to c \in (0,1] \text{ és} \\ & \liminf_{n,m\to\infty} a_{n2} > 1; \\ O\big(\frac{n-m}{n^{3/2}}\big), & \text{ha } \limsup_{n,m\to\infty} a_{n2} < 1. \end{cases}$$

A fentieket, illetve a 4.0.1. Tétel eredményét összehasonlítva látjuk, hogy az itt bevezetett diszkrét approximáció ugyanolyan, vagy jobb közelítését jelenti a várakozási időnek, mint a normális approximáció. Ráadásul a közelítés hibáját itt a Kolmogorov távolságnál sokkal erősebb teljes variációs távolságban mérjük.

Megjegyezzük, hogy a tételben a paraméterek úgy lettek megválasztva, hogy az összehasonlított eloszlások – $\pi_{\mu,a}$ és $\mathcal{D}(W_{n,m}+c)$ – első két momentuma megegyezzen.

A fejezet két alfejezetének eredményei publikálásra kerültek [15]-ben.

7. Poisson-Charlier sorfejtések

A dolgozat utolsó fejezetében a kupongyűjtő $\widetilde{W}_{n,m} = W_{n,m} - (n-m)$ eltolt várakozási idejét Poisson-Charlier előjeles mértékekkel közelítjük teljes variációs távolságban. A [4]-ben bevezetett karakterisztikus függvényeket használó módszert alkalmazzuk.

Legyen $\mu = \mathcal{D}(\widetilde{W}_{n,m}+c)$, ahol $c = \lfloor \mathbf{Var}\widetilde{W}_{n,m} - \mathbf{E}\widetilde{W}_{n,m} \rfloor$. Bevezetjük az

$$a_{n,j} := \sum_{k=m+1}^{n} \left(\frac{n-k}{k}\right)^j, \quad j = 1, 2, \ldots,$$

sorozatokat és a

$$h_R(w) = -\Big(a_{n,2} - \lfloor a_{n,2} \rfloor\Big) w + \Big(a_{n,2} - \lfloor a_{n,2} \rfloor\Big) \frac{w^2}{2} + \sum_{r=3}^{R} \Big(a_{n,r} + (-1)^{r+1} \lfloor a_{n,2} \rfloor\Big) \frac{w^r}{r},$$

valamint

$$H_R(w) = \begin{cases} \sum_{l=0}^{R} \frac{h_R^l(w)}{l!}, & \text{if } a_{n,2} > 1; \\ \sum_{l=0}^{3R-2} \frac{h_R^l(w)}{l!}, & \text{if } a_{n,2} < 1, \end{cases}$$

$w \in \mathbb{C}$, polinomokat. A μ valószínűségi mértéket a véges $\nu_R = \nu_R(\sigma_n^2, \widetilde{a}_{n,m}^{(1)}, \ldots, \widetilde{a}_{n,m}^{(R^2)})$ előjeles mértékkel közelítjük, melyet

$$\nu_R\{j\} = \mathrm{Po}(\sigma_n^2)\{j\} \left(1 + \sum_{r=1}^{S} (-1)^{r+1} \widetilde{a}_{n,m}^{(r)} C_r(j, \sigma_n^2)\right), \quad j \in \mathbb{N},$$

definiál, ahol $S = \deg(H_R(w))$, $\widetilde{a}_{n,m}^{(r)}$ az $(\mathrm{e}^{\mathrm{i}t} - 1)^r$ tag együtthatója a $H_R(\mathrm{e}^{\mathrm{i}t} - 1)$ polinomban, és $C_r(j, \sigma_n^2)$ az r-edik Charlier polinomot jelöli.

7.0.1. Tétel. *Feltesszük, hogy $a_{n,2} > 1$. Tetszőleges $R \geq 3$ esetén léteznek olyan R-től függő m_R és n_R küszöbszámok valamint C_R konstans, hogy ha $m \geq m_R$ és $n \geq n_R$, akkor*

$$\sup_{k \in \mathbb{Z}} |\mu\{k\} - \nu_R\{k\}| \leq C_R \left(\frac{1}{\sqrt{m}}\right)^R, \quad \text{ha } m \leq \frac{n}{2} - 1,$$

és

$$\sup_{k \in \mathbb{Z}} |\mu\{k\} - \nu_R\{k\}| \leq C_R \frac{(\sqrt{n})^{R-2}}{(n-m)^{R-1}}, \quad \text{ha } m \geq \frac{n}{2}.$$

7.0.2. Tétel. *Feltesszük, hogy $a_{n,2} < 1$. Tetszőleges $R \geq 3$ esetén létezik olyan R-től függő n_R küszöbszám és C_R konstans, hogy ha $n \geq n_R$, akkor*

$$\sup_{k \in \mathbb{Z}} |\mu\{k\} - \nu_R\{k\}| \leq C_R \frac{1}{(\sqrt{n})^R}.$$

A kapcsolódó teljes variációs távolságra vonatkozó eredmények kimondása előtt meghatározzuk az egymást követő R értékekhez tartozó közelítő mértékek különbségének teljes variációs normáját:

$$||\nu_{R+1} - \nu_R||_{TV} \leq \begin{cases} C_R \frac{1}{(\sqrt{m})^{R-1}}, & \text{a 7.0.1. Tétel esetén, ha } m \leq \frac{n}{2} - 1; \\ C_R \frac{(\sqrt{n})^{R-3}}{(n-m)^{R-2}}, & \text{a 7.0.1. Tétel esetén, ha } m \geq \frac{n}{2}; \\ C_R \frac{n-m}{(\sqrt{n})^{R+1}}, & \text{a 7.0.2. Tétel esetén.} \end{cases}$$

7.0.1. Következmény. *Minden olyan n és m esetén, melyre a 7.0.1. Tétel érvényes,*

$$d_{\mathrm{TV}}(\mu, \nu_R) \leq C_R \sigma_n \log \sigma_n \left(\frac{1}{\sqrt{m}}\right)^R, \quad \text{ha } m \leq \frac{n}{2} - 1,$$

és

$$d_{\mathrm{TV}}(\mu, \nu_R) \le C_R \frac{(\sqrt{n})^{R-3}}{(n-m)^{R-2}}, \quad ha\ m \ge \frac{n}{2};$$

továbbá, minden olyan n és m esetén, melyre a 7.0.2. Tétel érvényes,

$$d_{\mathrm{TV}}(\mu, \nu_R) \le C_R \frac{n-m}{(\sqrt{n})^{R+1}},$$

ahol C_R R-től függő pozitív konstans.

Ha összehasonlítjuk a 7.0.1. Következményt és a $||\nu_{R+1} - \nu_R||_{TV}$ normára kapott korlátokat, látjuk, hogy eredményeink nem optimálisak az $m \le n/2 - 1$ esetben abban az értelemben, hogy a ν_R-rel történő approximáció hibájának rendje nem egyezik meg az $(R+1)$-edik korrekciós tag $||\nu_{R+1} - \nu_R||_{TV}$ teljes variációs normájának rendjével. Érdekes nyitott probléma, hogy $d_{\mathrm{TV}}(\mu, \nu_R) \le C_R/(\sqrt{m})^{R-1}$ elérhető-e $m \le n/2 - 1$ esetén.

A dolgozat végén összehasonlítjuk az utolsó fejezet eredményeit a 6. fejezetben az összetett Poisson approximáció esetén kapottakkal.

Hivatkozások

[1] Banderier, C. and Dobrow, R. P., A Generalized Cover Time for Random Walks on Graphs, Proceedings of FPSAC'00, 2000.

[2] Barbour, A.D. and Cekanavicius, V, Total variation asymptotics for sums of independent integer random variables, *The Annals of Probability* **30** (2002), 509–545.

[3] Barbour, A.D. and Hall, P., On the rate of Poisson convergence, *Math. Proc. Cam. Phil. Soc.* **95** (1984), 473–480.

[4] Barbour, A. D., Kowalski, E., Nikeghbali, A., *Mod-discrete expansions*, *arXiv:0912.1886v1 [math.PR]*, 2009.

[5] Barbour, A.D. and Xia, A., Poisson Perturbations, *ESAIM Probab. and Statist.* **3** (1999), 131–150.

[6] Baum, L.E. and Billingsley, P., Asymptotic distributions for the coupon collector's problem, *Ann. Math. Statist.* **36** (1965), 1835–1839.

[7] Chihara, T. S., *An introduction to orthogonal polynomials*, Gordon and Breach, New York, 1978.

[8] Erdős, P. and Rényi, A., On a classical problem of probability theory, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **6** (1961), 215–220.

[9] Feller, W., *An Introduction to Probability Theory and its Applications*, John Wiley & Sons, 1968.

[10] Gnedenko, B. V. and Kolmogorov, A. N., *Limit Distributions for Sums of Independent Random Variables*, Addison-Wesley Publishing Company, Cambridge, Mass., 1954.

[11] Gut, A. and Holst, L., On the waiting time in a generalized roulette game, *Statistics & Probability Letters*, **2** (1984), 229–239.

[12] Holst, L., The general birthday problem, Proceedings of the sixth international seminar on Random graphs and probabilistic methods in combinatorics and computer science, John Wiley & Sons, 1995.

[13] Lindvall, T., *Lectures on the Coupling Method*, Dover Publications, 1992.

[14] Pólya, G., Eine Wahrscheinlichkeitsaufgabe zur Kundenwerbung, *Z. Angew. Math. Mech.*, **10** (1930), 96–97.

[15] Pósfai A., An extension of Mineka's coupling inequality, *Electronic Communications in Probability*, **14** (2009), 464–473.

[16] Pósfai, A., A supplement to the paper Poisson approximation in a Poisson limit theorem inspired by coupon collecting, *arXiv:0904.4924 [math.PR]*, 2009.

[17] Pósfai A., Poisson approximation in a Poisson limit theorem inspired by coupon collecting, *Journal of Applied Probability*, **46** (2009), 585–592.

[18] Pósfai, A., Rates of convergence for normal approximation in incomplete coupon collection, *Acta Scientiarum Mathematicarum (Szeged)* **73** (2007), 333–348.

[19] Pósfai, A. and Csörgő, S., Asymptotic approximations for coupon collectors, *Studia Scientiarum Mathematicarum Hungarica*, **46** (2009), 61–96.