Szerencsejátékok és a Big Data: valószínűségszámítás és statisztika gyakorlatban

Matematikai Intézet Nyílt napja, 2018. november 30.

Zempléni András

ELTE TTK Matematikai Intézet
Valószínűségelméleti és Statisztika Tanszék

Érdemes-e pénzt párna alatt tartani?

- Pénzünket olyan részvénybe fektetjük, amelynek értéke egy év alatt 50%-os eséllyel 1,9-szeresére nő és ugyanolyan eséllyel a felére csökken.
- Várható éves hozam: 0,5*0,9+0,5*(-0,5)=0,2=20%
- Az évek során pénzünk várható értéke végtelenhez tart.

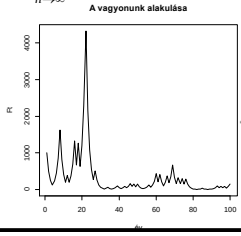
Érdemes-e pénzt párna alatt tartani? (folytatás)

$$S_n = X_1 \cdot \ldots \cdot X_n = e^{\ln(X_1)+\ldots+\ln(X_n)} = \left(e^{\frac{\ln(X_1)+\ldots+\ln(X_n)}{n}}\right)^n$$

$$\frac{\ln(X_1)+\ldots+\ln(X_n)}{n} \xrightarrow[n\to\infty]{} -1{,}114\% \Rightarrow$$

$$S_n \xrightarrow[n\to\infty]{} 0$$

Egy példa-futás:



1000 szimuláció eredménye

- Kiinduló tőke: 1000 Ft

Min.	1. Kv.	Medián	Átlag	3. Kv.	Max.
0.000e+00	1.000e+00	4.000e+01	2.84e+07	8.44e+03	2.015e+10

- Az esetek több, mint 66%-ában kevesebb lesz a pénzünk 1000 Ft-nál

A pénz fele párna alatt

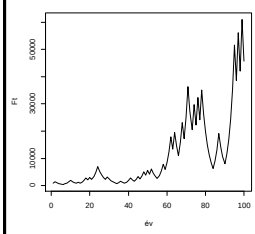
$$V_n = Y_1 \cdot \ldots \cdot Y_n = e^{\ln(Y_1)+\ldots+\ln(Y_n)} = \left(e^{\frac{\ln(Y_1)+\ldots+\ln(Y_n)}{n}}\right)^n$$

Várható éves hozam: 10%

$$\frac{\ln(Y_1)+\ldots+\ln(Y_n)}{n} \xrightarrow[n\to\infty]{} 1{,}821\% \Rightarrow$$

$$V_n \xrightarrow[n\to\infty]{} \infty$$

Eredmények



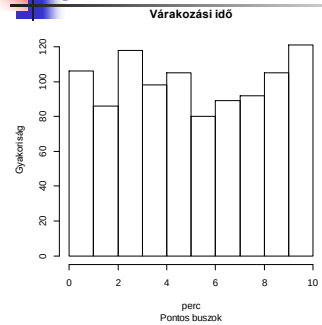
- 1000 szimuláció alapján:

Min.	1.kvartilis	Medián
0	6327	45720
Átlag	3.kvartilis	Max.
4692000	638700	$4{,}66*10^8$

Buszparadoxon

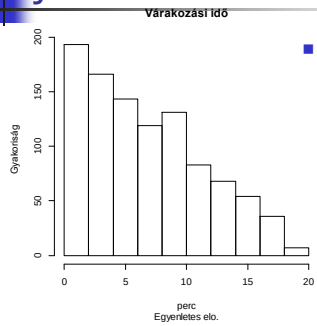
- Az autóbuszok átlagosan 10 percenként követik egymást. Egy buszmegállóba odamenve várhatóan hány percet kell várnunk a buszra?
- 5 percet, 10 percet, végtelen sokat?

Pontosan 10 percenként jönnek a buszok



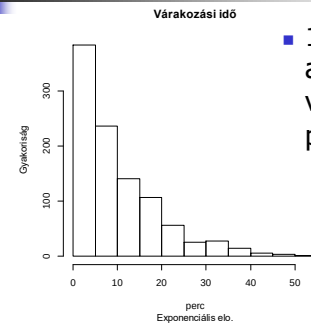
- 1000 szimuláció alapján az átlagos várakozási idő 5 perc

Átlagosan 10 percenként jönnek a buszok



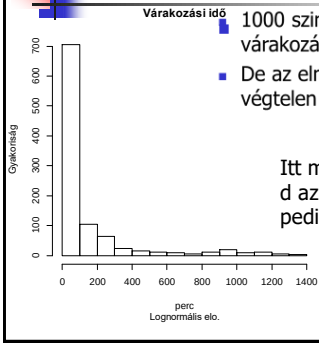
- 1000 szimuláció alapján az átlagos várakozási idő 6,7 perc

Átlagosan 10 percenként jönnek a buszok



- 1000 szimuláció alapján az átlagos várakozási idő 10 perc

Átlagosan 10 percenként jönnek a buszok



- 1000 szimuláció alapján az átlagos várakozási idő 145 perc
- De az elméleti várható érték akár végtelen sok is lehetne. A képlet:

$$m = \frac{1}{2}d\left(1 + \frac{s}{d^2}\right)$$

Itt m az átlagos várakozási időnk, d az átlagos követési időköz, s pedig a követési időköz varianciája

Elemzés, értékelés, döntéselőkészítés

- Utolsó példák már ilyenek
- Általában nem tudhatjuk a pontos valószínűségeket
- Konkrét megfigyelések alapján becslünk.
- Matematikai statisztika!
- Kérdések:
  - Mi a legjobb módszer?
  - Mennyire pontos/megbízható a becslésünk/döntésünk?

Statisztika

- Regresszió: X (ismert mennyiség) függvényével közelítjük Y értékét
- Példák:
  - Y: holnapi tőzsdeindex, X: mai (és megelőző napokon mért) értékek
  - Y: holnapi hőmérséklet, X: mai (és megelőző napokon mért) értékek
- Módszerek:
  - Lineáris regresszió
  - k legközelebbi szomszéd módszer

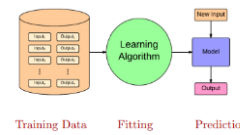
Paradoxonok

- Vigyázat: a kapcsolat léte még nem jelent ok-okozati viszonyt!
  - Cápatámadások száma és az eladott fagylaltok száma
  - Elfűtött tüzelőanyag és a szoba hőmérséklete

Modern fogalmak

- Gépi tanulás (az adatokból tanulnak az algoritmusok)
- Statisztikai tanulás (emellett még a bizonytalanságot is próbálja mérni)
- Adattudomány (mindenből egy kicsi…)
- Big data: nemcsak a nagysága miatt más, hanem a struktúrája miatt is:
  - Sok különböző adatstruktúra
  - Erősen hiányos adatbázisok
- A cél az összefüggések feltárása

Felügyelt tanulás



- Hagyományos statisztika: hosszú előkészítés, adattisztítás
- Modern megközelítés: nagy adat sok tulajdonsággal – gépi tanulás segít ezek közül a fontosakat kiválasztani
- Lényeg: az új adatokon működjön a módszer!

Mi is az a „big data”?

- Nincs matematikai definíció
- Nagy (terabyte-petabyte), változó, komplex (pl. szöveg-képek-video)
- A cél az összefüggések kinyerése
- Lehetséges módszerek:
  - Részek elemzése, összesítése
  - Véletlen részminták vizsgálata
  - Mesterséges intelligencia

A szimuláció (és a big data elemzés) eszköze: R

- Az R egy statisztikai programcsomag
- Nyílt forráskódú
- Szabadon letölthető: cran.r-project.org
- Magyar nyelven is hozzáférhető sok segédanyag
- A szimulációnál használt program: http://zempleni.elte.hu/kutej.txt

Ipari alkalmazások

- Bonyolult rendszerek monitorozása, a potenciális hibaforrások előzetes beazonosítása (pl. olajfúrótornyoknál)
- App a kagylótenyésztő farmereknek a tápanyag- és egyéb faktorok optimalizálására (ha már ez megvan, hasonlók sok helyen elképzelhetőek)
- Szállítócégek költségszámítása – jó előre, így könnyű optimalizálni: mit mikor, mivel érdemes szállítani

Játékok – a mesterséges intelligencia a nyerő

- AlphaGoZero: Úgy tanulta meg a go játékot, hogy nem kapott semmi külső segítséget. Csak a szabályokra épülve, 24 órai tanulás után legyőzte a korábbi világbajnok programot
- Hasonló eredményt ért el a sakkban is
- Fő eszköze a Google gépi tanulásra optimalizált chipje, a TPU (tensor processing unit)
- Ezek alapköve a Tensorflow, amivel neurális hálókat lehet konstruálni és tanítani

Források

- Wang et. al: Statistical methods and computing for big data (2016), U.S. Department of Health and Human Services
- Boulton: 10 data analytics success stories: An inside-look (2018) https://www.cio.com/article/3221621/analytics/6-data-analytics-success-stories-an-inside-look.html