

MATEMATIKA ÉS STATISZTIKA DATA SCIENCE-HEZ



VALÓSZÍNŰSÉGSZÁMÍTÁS II. VALÓSZÍNŰSÉGI VÁLTOZÓK, NEVEZETES VALÓSZÍNŰSÉGI ELOSZLÁSOK, CHT

VALÓSZÍNŰSÉGI VÁLTOZÓK



Eddig **események valószínűségének kiszámításával** foglalkoztunk: kísérletek kimenetelei \longrightarrow szám

Ezt a minden kimenetelhez egy számot hozzárendelő függvényt nevezzük **valószínűségi változónak**.

DISZKRÉT VALÓSZÍNŰSÉGI VÁLTOZÓ

- ▶ értékkészlete véges vagy
- ▶ megszámlálhatóan végtelen számosságú

X_1 = egy adott felhasználó reakciója egy termék A/B tesztverziójára

$$X = \begin{cases} 1, & \text{ha konvertál (pl. vásárol)} \\ 0, & \text{ha nem konvertál} \end{cases}$$

X_2 = egy felhasználó hányszor látogat vissza a honlapra egy adott időablakon belül

$$X_2 \in \{ 0; 1; 2; 3; 4; \dots \}$$

(elméletileg végtelen, de megszámlálható)

FOLYTONOS VALÓSZÍNŰSÉGI VÁLTOZÓ

- ▶ értékkészlete végtelen számosságú

X_3 = a következő géphiba bekövetkezéséig eltelt idő (általában órákban)

$$X_3 \in [0; \infty [$$

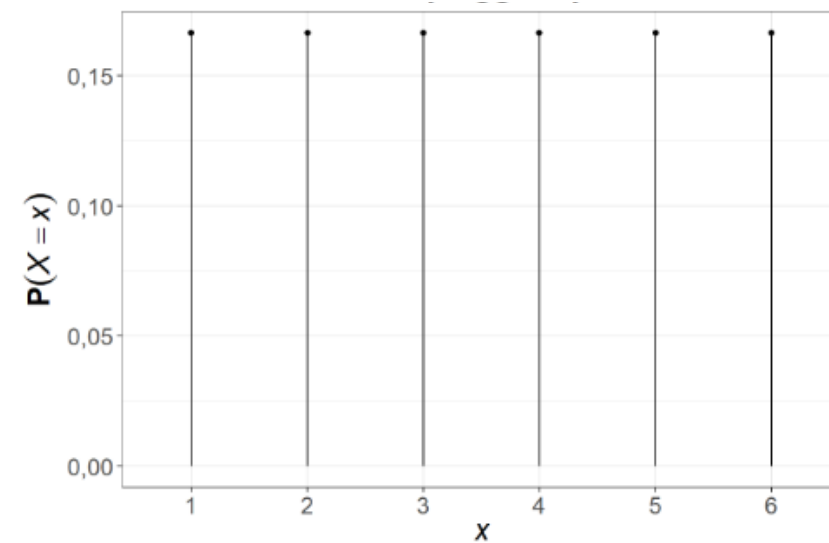
X_4 = egy ML modell által becsült valószínűség

$$X_4 \in [0; 1]$$

VALÓSZÍNŰSÉGI VÁLTOZÓK JELLEMZÉSE FÜGGVÉNYEK SEGÍTSÉGÉVEL



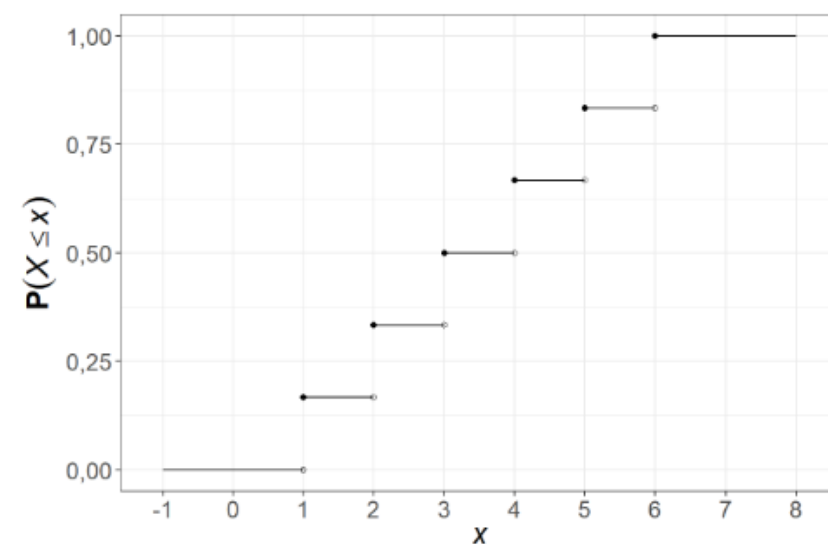
DISZKRÉT VV SÚLYFÜGGVÉNYE



$$\mathbf{P}(X = x_k) = p_k$$

ami egy tetszőleges valós számhoz az érték bekövetkezési valószínűségét rendeli

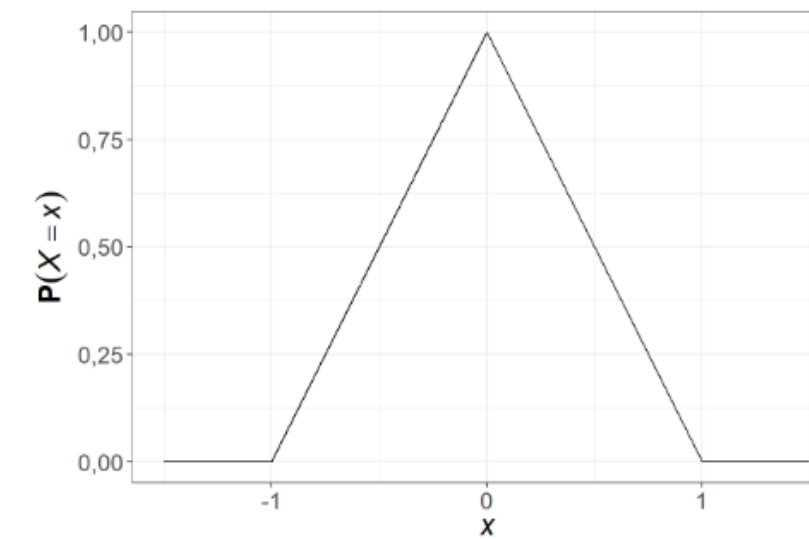
DISZKRÉT VV ELOSZLÁSFÜGGVÉNYE



$$F(x) = \mathbf{P}(X \leq x)$$

ami egy tetszőleges valós számhoz az $X \leq x$ esemény bekövetkezési valószínűségét rendeli.

FOLYTONOS VV SŰRŰSÉGFÜGGVÉNYE

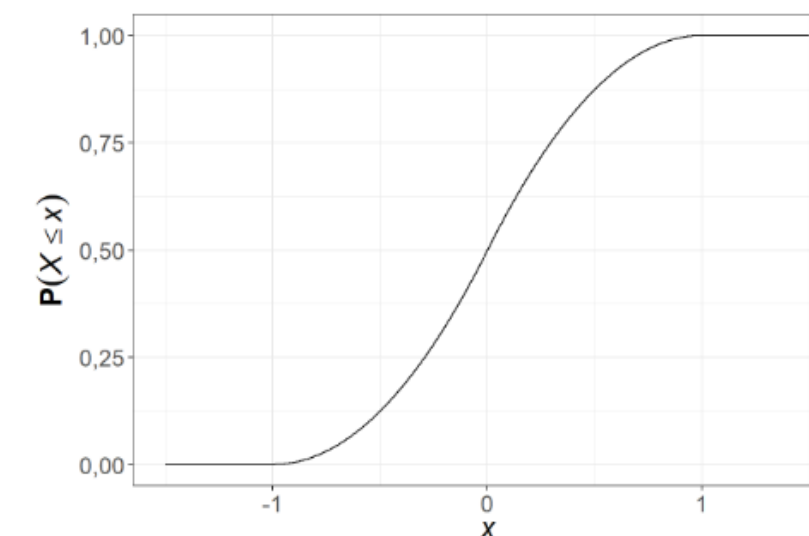


$$F(x) = \int_{-\infty}^x f(u) du$$

$$f(x) = F'(x)$$

ami megmutatja, hogy az adott x érték környezetében mekkora a valószínűségi sűrűség (azaz, hogy milyen intenzíven fordulnak elő értékek körülötte)

FOLYTONOS VV ELOSZLÁSFÜGGVÉNYE



$$F(x) = \mathbf{P}(X \leq x)$$

ami egy tetszőleges valós számhoz az $X \leq x$ esemény bekövetkezési valószínűségét rendeli.

VALÓSZÍNŰSÉGI VÁLTOZÓK JELLEMZÉSE MOMENTUMOK SEGÍTSÉGÉVEL



DISZKRÉT VV VÁRHATÓ ÉRTÉKE

$$\mathbf{E}(X) = \sum_k x_k \cdot p_k$$

A **várható érték** megmutatja, hogy átlagosan milyen értéket várunk a valószínűségi változótól, ha végtelen sok megfigyelést végeznénk.

$$X \in \{1, 2, 3, 4, 5, 6\}, \quad P(X = x_i) = \frac{1}{6}$$

$$\mathbb{E}[X] = \sum_{i=1}^6 x_i \cdot \frac{1}{6} = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$$

DISZKRÉT VV VARIANCIÁJA ÉS SZÓRÁSA

$$\mathbf{D}^2(X) = \sum_k (x_k - \mathbf{E}(X))^2 \cdot p_k$$

A **variancia** megmutatja, hogy a valószínűségi változó értékei átlagosan mennyire térnek el a várható értéktől.

A **szórás** ennek a négyzetgyökét jelenti, így ugyanabban a mértékegységben van, mint maga a változó.

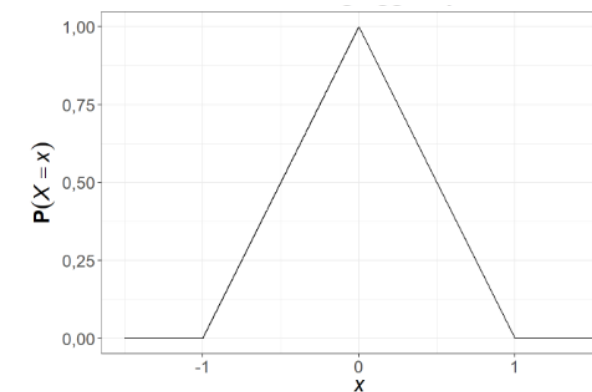
$$X \in \{1, 2, 3, 4, 5, 6\}, \quad P(X = x_i) = \frac{1}{6}, \quad E[X] = 3.5$$

$$D^2(X) = \sum_{i=1}^6 (x_i - 3.5)^2 \cdot \frac{1}{6} = \frac{1}{6}(6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25) = 2.9167$$

$$D(X) = \sqrt{2.9167} \approx 1.71$$

FOLYTONOS VV VÁRHATÓ ÉRTÉKE

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$



$$f(x) = \begin{cases} 1 - |x|, & \text{ha } -1 \leq x \leq 1, \\ 0, & \text{egyébként.} \end{cases}$$

$$E[X] = 0.$$

$$E[X] = \int_{-1}^1 x(1 - |x|) dx.$$

FOLYTONOS VV VARIANCIÁJA ÉS SZÓRÁSA

$$\mathbf{D}^2(X) = \int_{-\infty}^{\infty} (x - \mathbf{E}(X))^2 f(x) dx$$

$$f(x) = 1 - |x|, \quad -1 \leq x \leq 1, \quad E[X] = 0$$

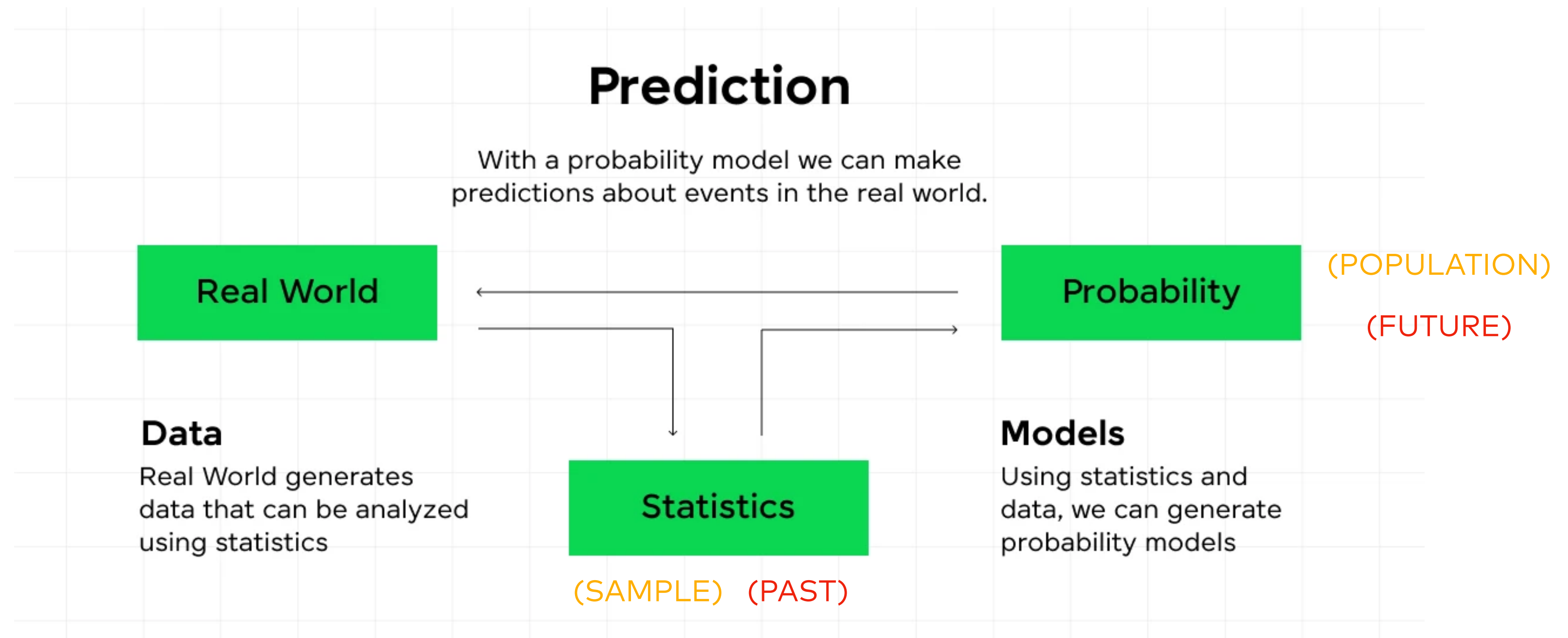
$$D^2(X) = 2 \int_0^1 x^2(1 - x) dx = 2 \left[\frac{x^3}{3} - \frac{x^4}{4} \right]_0^1 = 2 \left(\frac{1}{3} - \frac{1}{4} \right) = \frac{1}{6}$$

$$D(X) = \sqrt{\frac{1}{6}} \approx 0.41$$

STATISZTIKAI VS VALÓSZÍNŰSÉGI VÁLTOZÓK

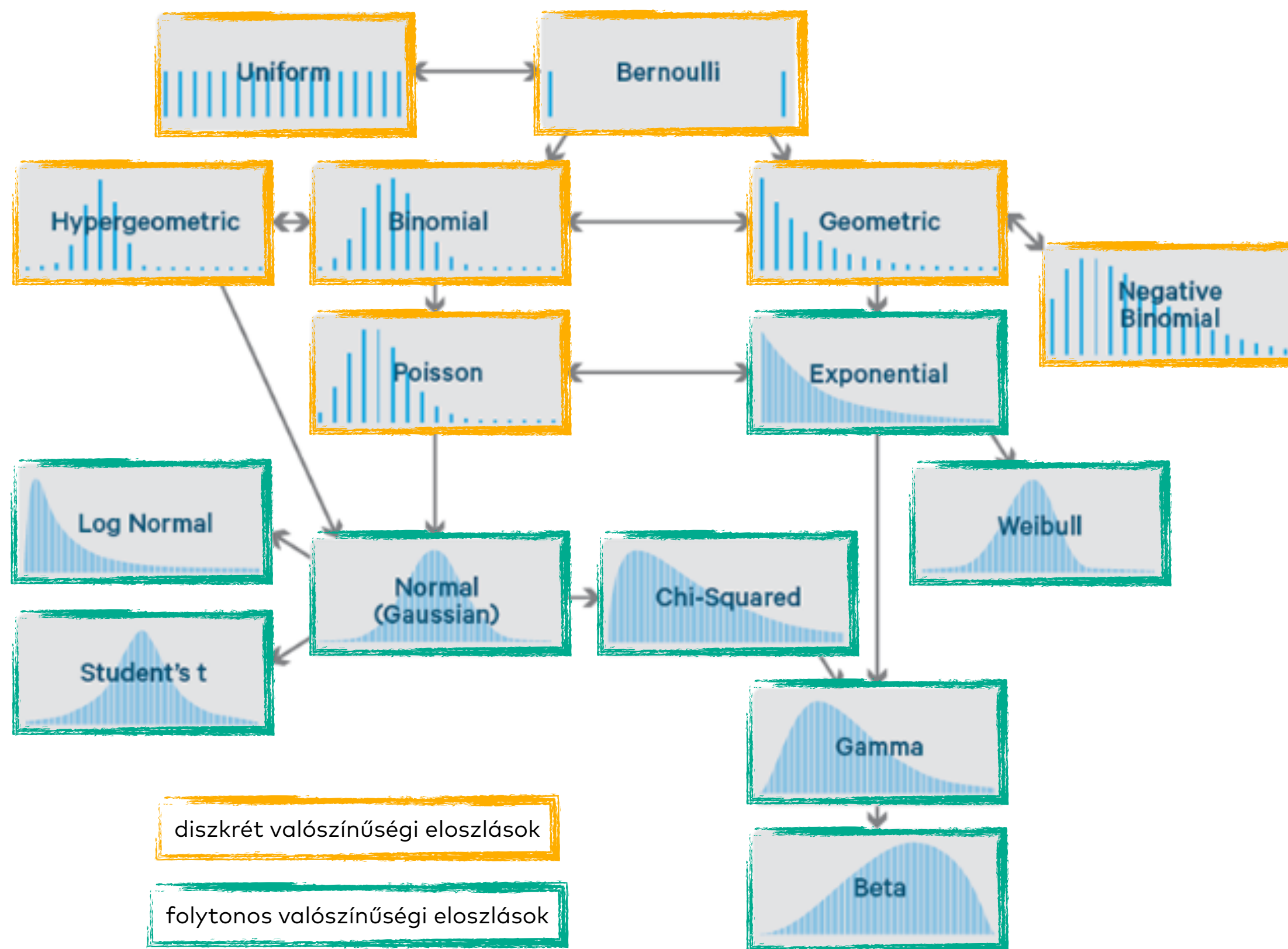


STATISZTIKA VS VALÓSZÍNŰSÉG A DS-BEN



[kép forrása: <https://www.guvi.in/blog/probability-and-statistics-for-data-science/>]

NEVEZETES VALÓSZÍNŰSÉGI ELOSZLÁSOK



A valószínűségi modelleknek szüksége van **a jelenségeket jellemző törvényszerűségeket összegző mintázatokra.**

Nevezetes valószínűségi eloszlások

A modellek ezeknek az eloszlásoknak a jellemzőit használják fel:

- ▶ **várható értéket** → átlagos viselkedés becsléséhez,
- ▶ **szórást** → bizonytalanság méréséhez,
- ▶ **eloszlásalakot** → valószínűségek számításához és döntési küszöbökhöz

NEVEZETES DISZKRÉT ELOSZLÁSOK + DS



Eloszlás	Fő jellemzők	Használat DS-ben	Kapcsolódó statisztikai és/vagy ML modellek	Hogyan használja a modell az eloszlás jellemzőit?
Bernoulli	Egyetlen 0/1 kimenet (siker vagy kudarc)	Bináris események, minőségellenőrzés, konverzió	Logisztikus regresszió, Naive Bayes, neurális hálók	A modell a siker valószínűségét tanulja → a veszteség-függvény (log loss) ezen Bernoulli valószínűségekre épül
Binomiális	Sikeres események száma n független kísérletben	A/B teszt, konverziók, mintabeli arányok	Logisztikus regresszió, Z-teszt arányokra	A modell a mintaarány szórását tanulja → konfidenciaintervallumokat és p-értékeket számít
Poisson	Események száma adott idő/tér intervallumban	Kattintások, hibák, érkezések modellezése	Poisson-regresszió, count modellek, időbeli előrejelzés	A modell a várható eseményszámot (λ) tanulja → a log-likelihood függvény a Poisson valószínűségekre épül
Geometriai	Hány próbálkozás kell az első sikerig	Kísérletek, próbálkozások elemzése (ügyfélsiker, CRM)	Szekvenciális modellek, RL exploration	A modell a siker valószínűségét becsüli, és az alapján súlyozza, milyen gyorsan várható az első pozitív esemény.

NEVEZETES FOLYTONOS ELOSZLÁSOK + DS



Eloszlás	Fő jellemzők	Használat DS-ben	Kapcsolódó statisztikai és/vagy ML modellek	Hogyan használja a modell az eloszlás jellemzőit?
Normal	Szimmetrikus, haranggörbe alakú; az értékek az átlag körül sűrűsödnek	Feature scaling, outlier-detekció, hibamodellezés	Lineáris regresszió, ANOVA, t-teszt, z-teszt, PCA	A modell feltételezi, hogy a hibák normál eloszlásúak → az átlag és szórás alapján határozza meg a valószínűséget, konfidenciát
Log-normál	A logaritmusuk szerint normális eloszlású mennyiségek	Jövedelmek, árfolyamok, növekedések modellezése	Log-transzformált regresszió, árfolyam-modellek	A modell log térben normál eloszlást feltételez → így a multiplikatív hatásokat additívvá alakítja
Exponenciális	Az események közötti várakozási idő eloszlása	Időközök modellezése (pl. ügyfélérkezés, hibák közti idő)	Survival analysis, hazard modellek	A modell a várható érték reciproka (λ) alapján becsüli az esemény bekövetkezésének intenzitását (rate parameter)
Khi-négyzet	Kategóriák közti eltérést méri; a négyzetösszegek eloszlása	Kategórikus adatok függetlenségvizsgálata	Khi-négyzet próba, döntési fák splitting-kritériuma	A modell az eltérések négyzetösszegét méri → a Khi-négyzet eloszlás alapján dönti el, hogy eltérés szignifikáns-e.

VALÓSZÍNŰSÉGI ELOSZLÁSOKRÓL BŐVEBBEN...



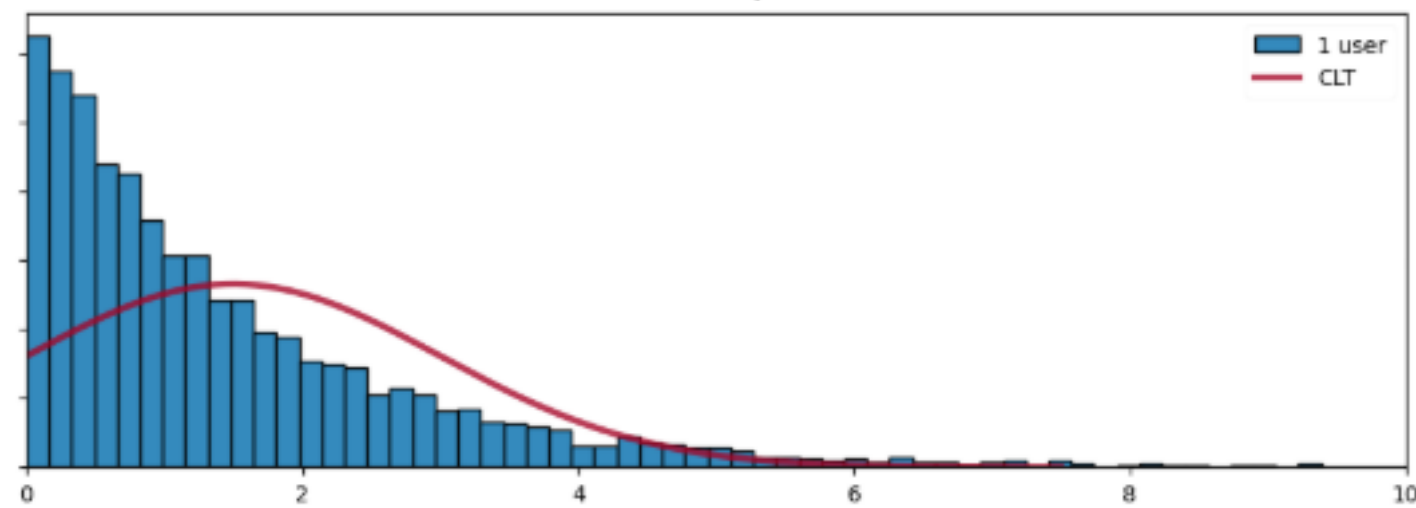
<https://www.geeksforgeeks.org/data-science/probability-data-distributions-in-data-science/>

CENTRÁLIS HATÁRELOSZLÁS TÉTEL

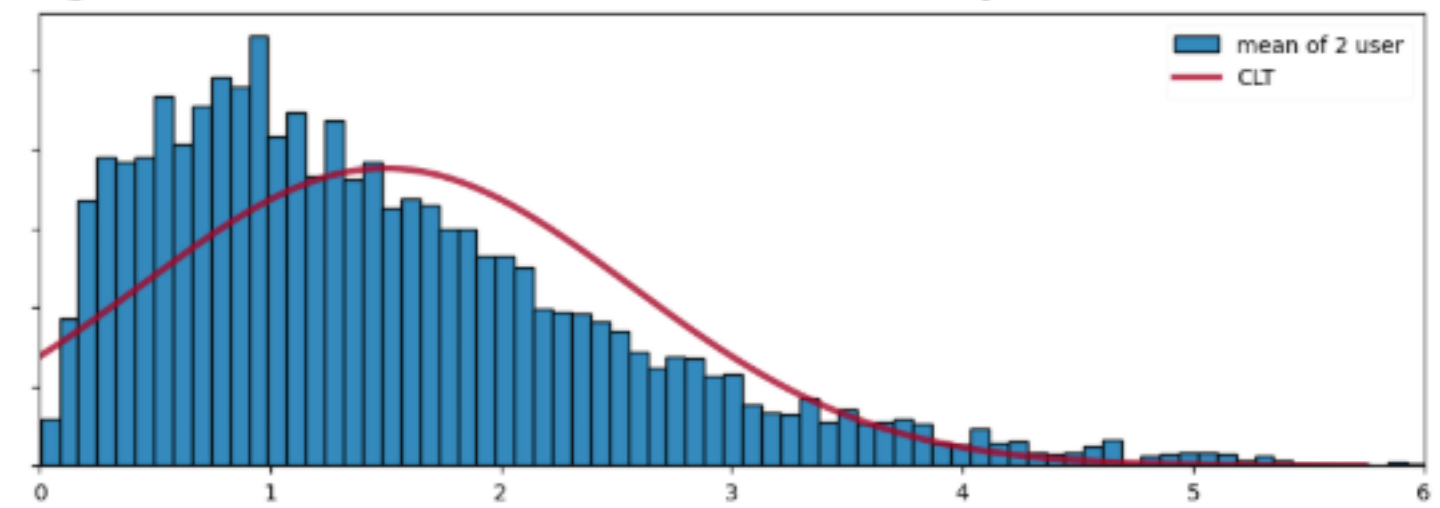


Miért és hogyan lesz végül minden "átlag" normális eloszlású? És miért jó ez nekünk?

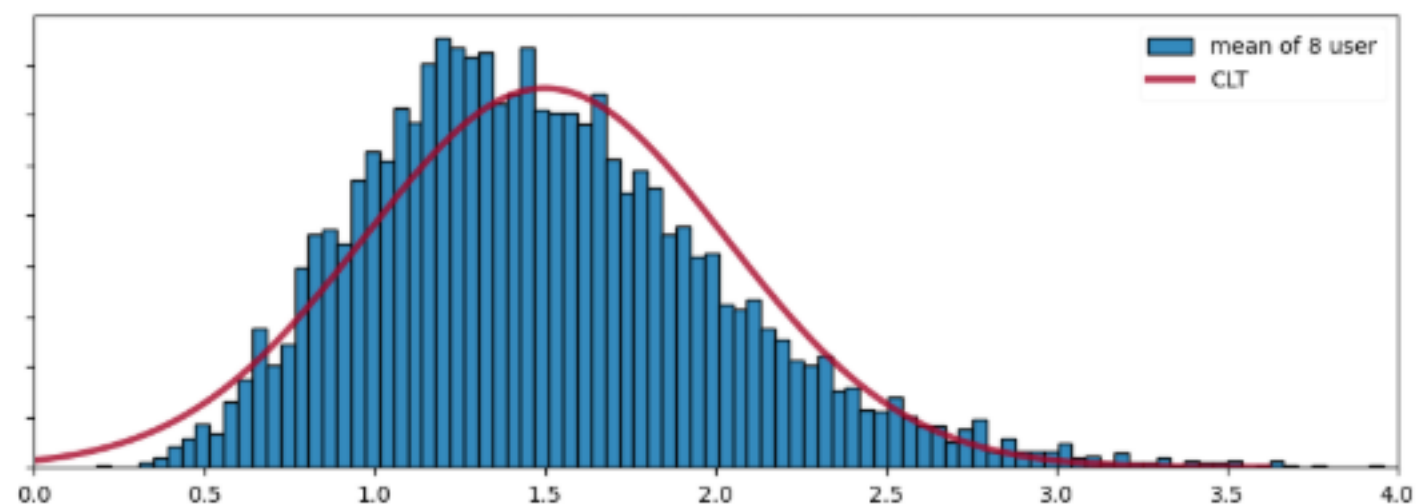
Legyen az alap szituáció, amit vizsgálunk felhasználók webes eseményei között eltelt idő.
Tegyük fel, hogy minden felhasználótól van 5000 adatpontunk.



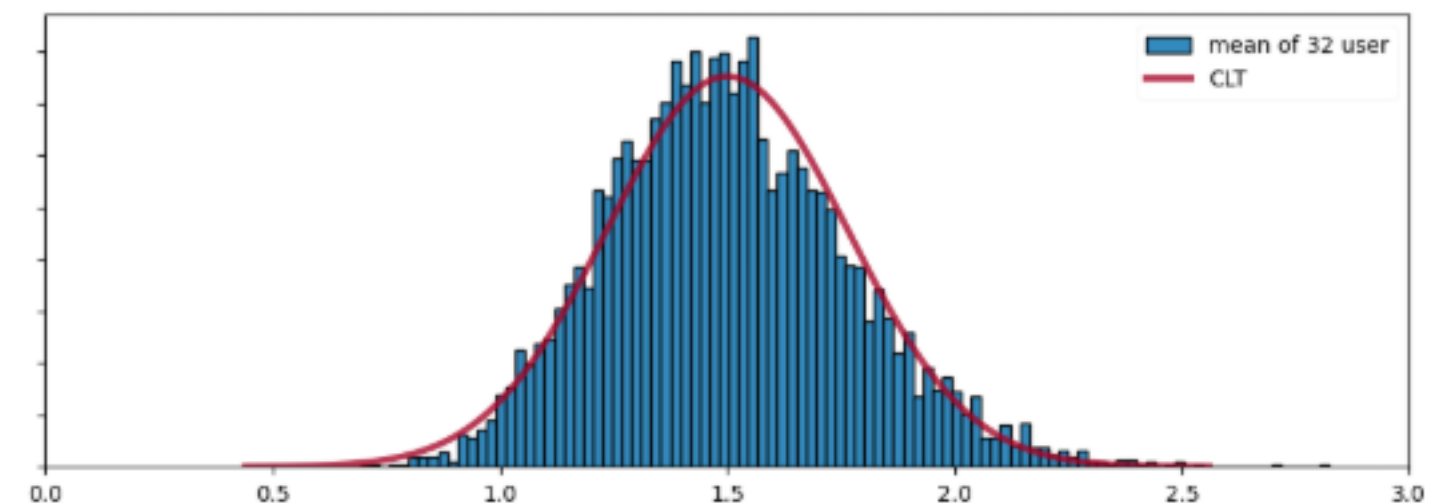
egyetlen felhasználó két egymást követő webes eseménye (pl. két kattintás) között eltelt idő



2 felhasználó adataiból számolt **átlagok eloszlása**



8 felhasználó adataiból számolt **átlagok eloszlása**



32 felhasználó adataiból számolt **átlagok eloszlása**

CENTRÁLIS HATÁRELOSZLÁS TÉTEL



A **Centrális Határeloszlás Tétel** (**CHT**, angolul **CLT**) azt mondja ki, hogy

- ▶ ha egy populációból sok véletlenszerű mintát veszünk,
- ▶ és mindegyik minta átlagát kiszámítjuk, akkor elegendően nagy n esetén
- ▶ ezeknek **az átlagoknak az eloszlása közelít a normális eloszláshoz**
- ▶ függetlenül attól, hogy az eredeti adatok milyen eloszlásúak voltak

X_1, X_2, \dots, X_n független, azonos eloszlású változók,

$$E[X_i] = \mu, \quad D[X_i] = \sigma^2$$

$$\Rightarrow \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

ÉS

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \rightarrow \mathcal{N}(0, 1)$$

CENTRÁLIS HATÁRELOSZLÁS TÉTEL + DS



ezen alapulnak **a ML modelljeink becslései**

- ▶ LINREG — sok kis hiba összegzése után a hibák eloszlása normál eloszlású lesz;
- ▶ BAGGING alapú fa modellek — a fák predikcióinak átlaga normális hibaeloszlású lesz;
- ▶ BOOSTING alapú fa modellek — a fák egymás hibáiból tanulnak, az összegezett hiba ismét sok független komponensből áll → ezért közel normális a végső residual

ezen alapulnak **a ML modellek értékelő metrikái**

- ▶ a regressziós modelleknél a predikciós hibák (pl. RMSE, MAE) átlagokból származnak, tehát CHT szerint normálisnak tekinthetők
- ▶ a klasszifikációs metrikák (accuracy, F1, AUC stb.) valójában mintákból számított átlagok vagy arányok — a CHT biztosítja, hogy ezek eloszlása közel normális legyen

ezen alapul **a Cross-validation technika**

- ▶ cross-validation során minden fold egy független mintán mért metrika
- ▶ a CHT miatt ezek az értékek normális eloszlást követnek, így megadhatjuk az átlagos teljesítményt \pm szórással
- ▶ a CHT az oka annak, hogy az átlagolt CV-score „megbízható” teljesítménymutató, és hogy így statisztikai alapon tudunk „best modell”-t választani

ezen alapulnak **a hipotézis vizsgálataink**
(a p-értékek és konfidencia intervallumok meghatározása)

- ▶ mert ezek mind abból indulnak ki, hogy a mintaátlag eloszlása normális — ez alapján tudjuk meghatározni, hogy egy becsült érték mennyire tér el a várttól, és mekkora a bizonytalanság a mintánk körül.

GYAKORLÓ FELADAT



Önállóan megoldandó kódolós feladat az érintett témakörökből:

[4_alkalom_gyakorlo_feladat.ipynb](#)