

# Vizuális adatelemzés

Kritikus Rendszerek Kutatócsoport

2021

## Tartalomjegyzék

<b>1. Kiegészítő anyagok*</b>	<b>2</b>	1.3. Kísérlettervezés . . . . .	<b>3</b>
1.1. Valószínűségszámítási alapfogalmak . . . . .	2	<b>Irodalomjegyzék</b>	<b>4</b>
1.2. Statisztikai alapfogalmak	2	<b>Tárgymutató</b>	<b>4</b>

## Bevezetés

Jegyzetként kérjük használják az „Intelligens adatelemzés” c. könyv 5. fejezetét (*Vizuális analízis*). Elérhető a <https://ftsrg.mit.bme.hu/remo-jegyzet/vizualis-adatelemzes-konyvfejezet.pdf> címen, ill. ingyenesen megvásárolható a [http://www.interkonyv.hu/konyvek/antal\\_peter\\_intelligens\\_adatelemzes](http://www.interkonyv.hu/konyvek/antal_peter_intelligens_adatelemzes) címen.

# 1. Kiegészítő anyagok\*

## 1.1. Valószínűségszámítási alapfogalmak

### Definíció.

- Valószínűségi változó (random variable):  $X$
- Várható érték, átlag (expected value, average, mean):

$$\mu = \mathbb{E}X = \sum_{i=1}^n p_i x_i$$

- Szórásnégyzet (variance):

$$\sigma^2 = \mathbb{E}(X - \mu)^2 = \sum_{i=1}^n p_i (x_i - \mu)^2$$

- Szórás (standard deviation):

$$\sigma = \sqrt{\mathbb{E}(X - \mu)^2} = \sqrt{\sum_{i=1}^n p_i (x_i - \mu)^2}$$

## 1.2. Statisztikai alapfogalmak

### Definíció.

- Minta (sample): megfigyelések (observation) halmaza,  $t$  darab,  $x_1, \dots, x_t$
- Tapasztalati átlag (sample mean):

$$m = \bar{x} = \frac{x_1 + \dots + x_t}{t}$$

- Korrigált tapasztalati szórás (unbiased sample standard deviation):

$$s = \sqrt{\frac{(x_1 - m)^2 + \dots + (x_t - m)^2}{t - 1}} = \sqrt{\frac{\sum_{i=1}^t (x_i - m)^2}{t - 1}}$$

Figyeljük meg, hogy a korrigált tapasztalati értékeknél  $t$  helyett  $(t - 1)$ -gyel osztunk. Ennek oka, hogy  $t$ -vel osztva a kapott érték általában alábecsli a teljes populáció szórását. Belátható, hogy  $(t - 1)$ -gyel osztva a valódi szórást jobban közelítő értéket kapunk. Ezt nevezzük Bessel-féle korrekciónak ([https://en.wikipedia.org/wiki/Bessel's\\_correction](https://en.wikipedia.org/wiki/Bessel's_correction)).

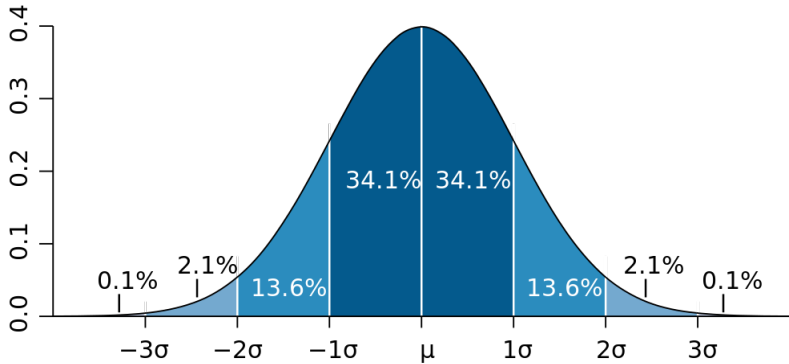
### 1.3. Kísérlettervezés

A *centrális határeloszlás-tétel*ből (*central limit theorem*) következik, hogy tetszőleges eloszlású jellemző (véges  $m$  várható értékkel és  $s$  szórással) tapasztalati átlaga  $t \rightarrow \infty$  esetén normális eloszlású,  $\mu = m$  várható értékkel és  $\sigma = \frac{s}{\sqrt{t}}$  szórással.

Ökölszabály: ismert szórásnál  $t > 30$ , ismeretlen szórásnál  $t > 100$  után kezd elfogadható lenni a közelítés.

A normális eloszlású változó

- az esetek 68%-ában legfeljebb  $1\sigma$  messze kerül  $\mu$ -tól,
- az esetek 95%-ában legfeljebb  $2\sigma$  messze kerül  $\mu$ -tól,
- az esetek 99,7%-ában legfeljebb  $3\sigma$  messze kerül  $\mu$ -tól.



1. ábra. Konfidenciaintervallumok

# Hivatkozások

## Tárgymutató

**átlag** average, mean 2

**CLT** centrális határeloszlás tétele; central limit theorem 3

**korrigált tapasztalati szórás** unbiased sample standard deviation 2

**megfigyelés** observation 2

**minta** sample 2

**szórás** standard deviation 2

**szórásnégyzet** variance 2

**tapasztalati átlag** sample mean 2

**valószínűségi változó** random variable  
2

**várható érték** expected value 2