

Statisztika a gyakorlatban

Ruzsa Zoltán
ruzsa.zoltan@emk.semmelweis.hu





A magyarok fele olyan településen él, ahol a jövedelem nem éri el az országos átlagot

VÁLASZTÁSI FÖLDRAJZ GAZDASÁG 2025. január 24. – 04:58



🚩 Legfontosabb



Mi várható a Kennedy-aktáktól, amikből Trump szerint minden kiderül?

KÜLFÖLD

Tematika

- Leíró statisztika (átlag, medián, módusz, szórás...)
- Paraméterbecslés (mennyi?)
- Hipotézisvizsgálat (igaz vagy nem?)
- Regressziószámítás, modellezés (hogy befolyásol?)



Tematika

- ~~Leíró statisztika (átlag, median, módsz, szórás...)~~
- Paraméterbecslés (mennyi?)
- Hipotézisvizsgálat (igaz vagy nem?)
- Regressziószámítás, modellezés (hogy befolyásol?)



Eszköz (R)

1) R + R studio saját gépen

vagy

2) Google colab: <https://colab.research.google.com/>





Barnóczki Brigitta



Másolás



2082

A Republikon Intézet legfrissebb [kutatása szerint](#) januárban hibahatáron belül tovább nőtt a Tisza Párt előnye a Fidesz–KDNP-vel szemben, a teljes népesség körében 3 százalékponttal vezet Magyar Péter pártja, 26 százalékpont a támogatottsága a Fidesz–KDNP 23-as eredményével szemben. Az MSZP-közeli intézet reprezentatív kutatása szerint a pártválasztók körében 37 százalék szavazna a Tiszára, 32 pedig a Fidesz–KDNP-re.

Módszertan: A kutatás 1000 fő telefonos megkérdezésével készült 2025. január 14–22. között. A kutatás nem, életkor, iskolai végzettség és településtípus szerint reprezentatív az ország felnőtt lakosságára. A hibahatár: +/- 3,2%.

Az intézet oldalán is elérhető [kutatás](#) szerint a vizsgált időszakban 39-32 volt az állás a biztos szavazó pártválasztók körében, ami az intézet szerint azt jelenti, hogy a Tisza magabiztosan vezet, de a növekedési üteme lassult. Az intézet azt

telex

Kövess minket Faceboo

Követem!

Legfontosabb



KÜLFÖLD

Mire a tűzoltó a negyedik és emeletek már emberek a fel színteken sik



A modern kor nyolcvan éve



Barnóczki Brigitta



Másolás



2082

A Republikon Intézet legfrissebb [kutatása szerint](#) januárban hibahatáron belül tovább nőtt a Tisza Párt előnye a Fidesz-KDNP-vel szemben, a teljes népesség körében 3 százalékponttal vezet Magyar Péter pártja, **26 százalékpont a támogatottsága a Fidesz-KDNP 23-as eredményével szemben.** Az MSZP-közeli intézet reprezentatív kutatása szerint a pártválasztók körében 37 százalék szavazna a Tiszára, 32 pedig a Fidesz-KDNP-re.

Módszertan: A kutatás 1000 fő telefonos megkérdezésével készült 2025. január 14–22. között. A kutatás nem, életkor, iskolai végzettség és településtípus szerint reprezentatív az ország felnőtt lakosságára. A hibahatár: +/- 3,2%.

Az intézet oldalán is elérhető [kutatás](#) szerint a vizsgált időszakban 39-32 volt az állás a biztos szavazó pártválasztók körében, ami az intézet szerint azt jelenti, hogy a Tisza magabiztosan vezet, de a növekedési üteme lassult. Az intézet azt

telex

Kövess minket Faceboo

Követem!

Legfontosabb

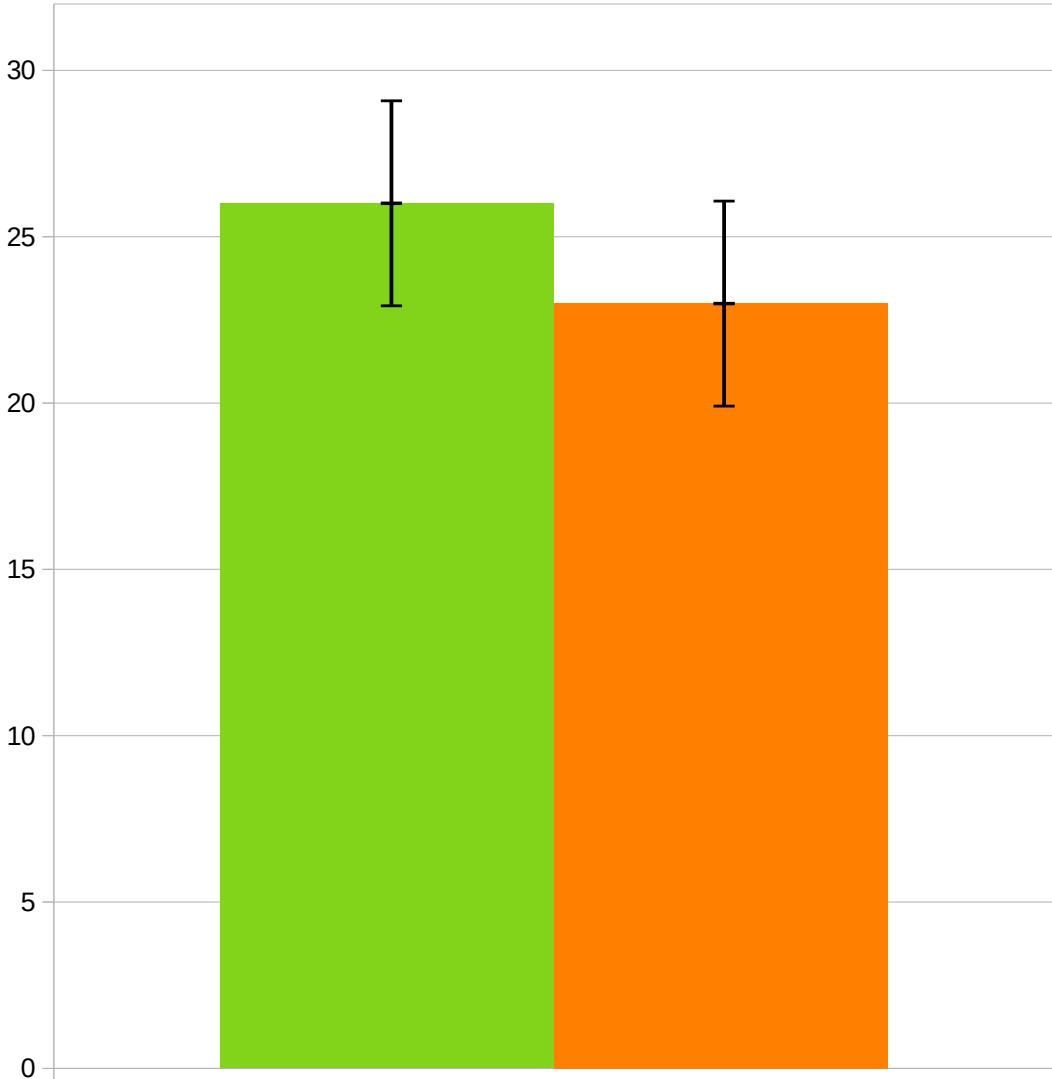


KÜLFÖLD

Mire a tűzoltó a negyedik és emeletek már emberek a fel színteken sik



A modern kor nyolcvan éve



Tisza: 26%

Fidesz: 23%

hibahatár: +/- 3.2%

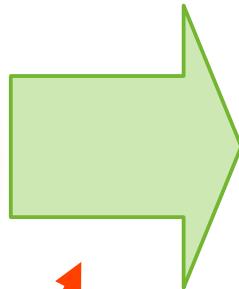
Valóság – matematikai modell – statisztika



Valóság – matematikai modell – statisztika



Módszertani hiba



8.2M szavazópolgár

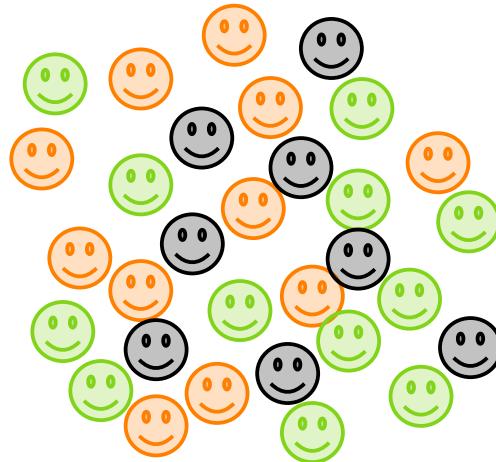


Mintavételi hiba



1000 fő

Matematikai modell – statisztika



8.2M ember

- f -ed része Fideszre szavaz
- t -ed része Tiszára
- $(1-f-t)$ -ed része egyéb

(f és t nem ismert paraméterek)



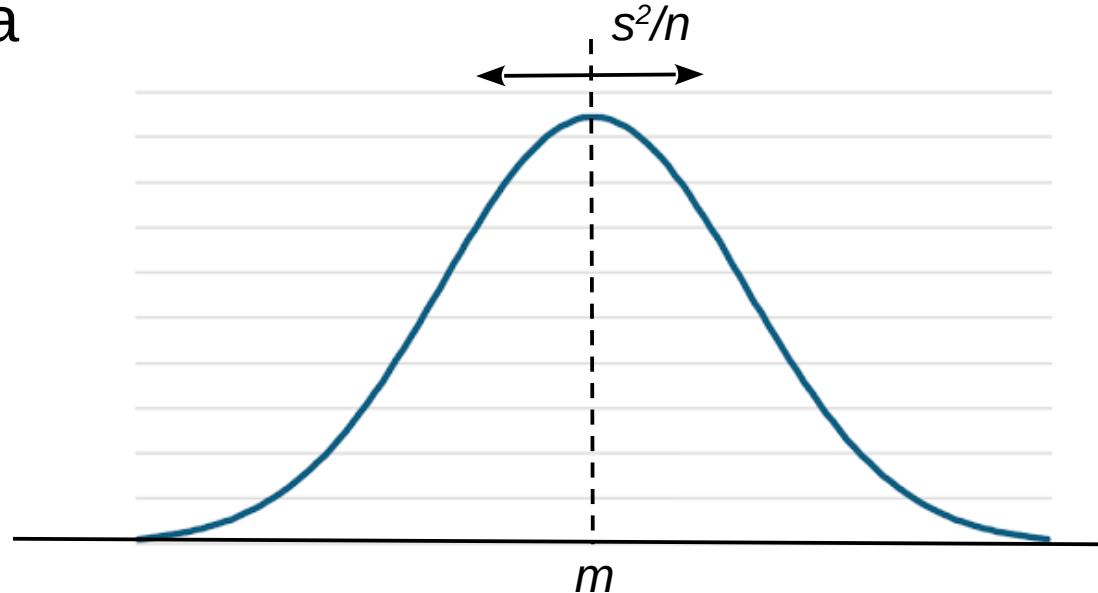
1000 fős minta

- tapasztalati becslés f -re
- tapasztalati becslés t -re

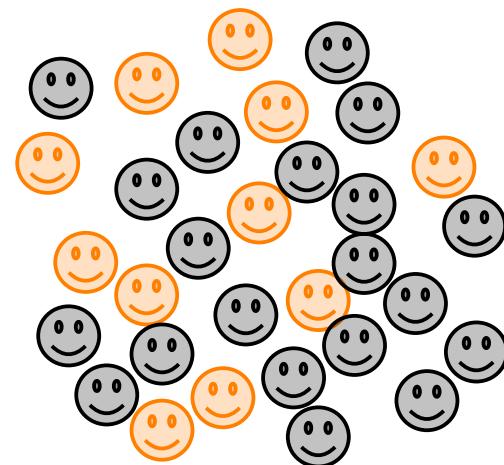
Centrális Határeloszlás Tétel

Bármilyen eloszlású valószínűségi változóból vett független minta átlaga közelítőleg normális eloszlású: $N(m, s^2/n)$, ahol

- m az eredeti változó várható értéke
- s az eredeti változó szórása
- n a minta elemszáma

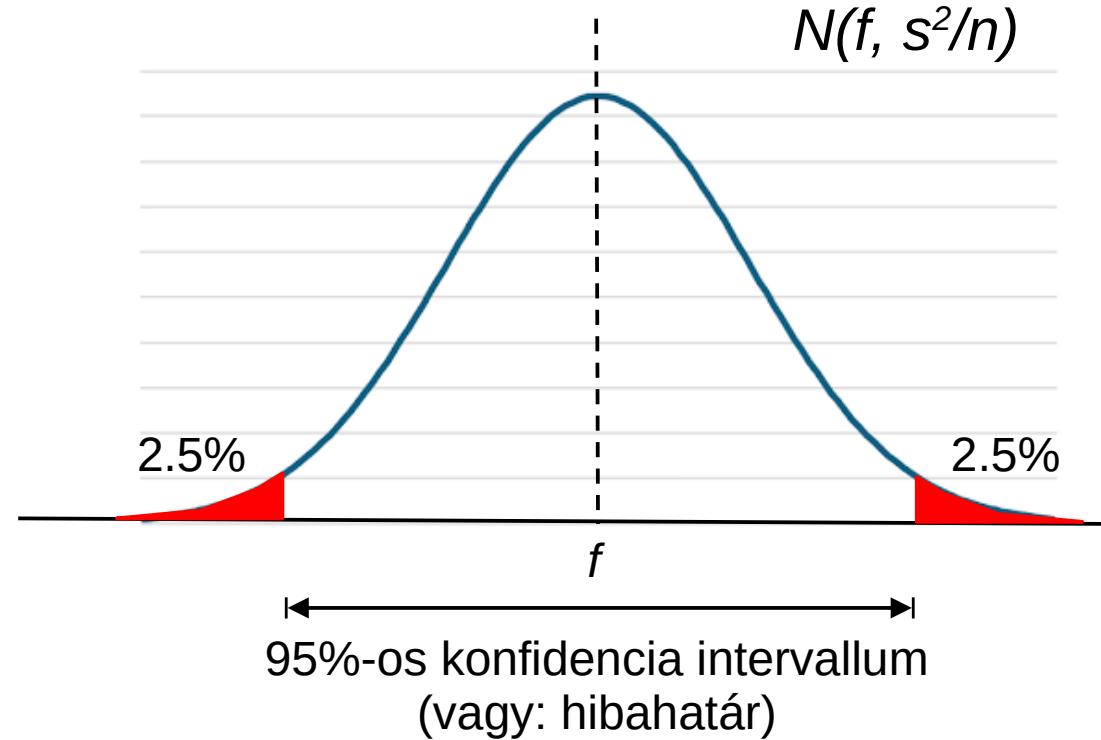


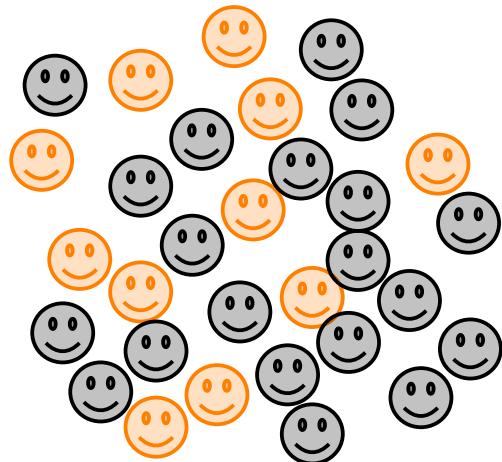
Konfidenciaszint → hibahatár



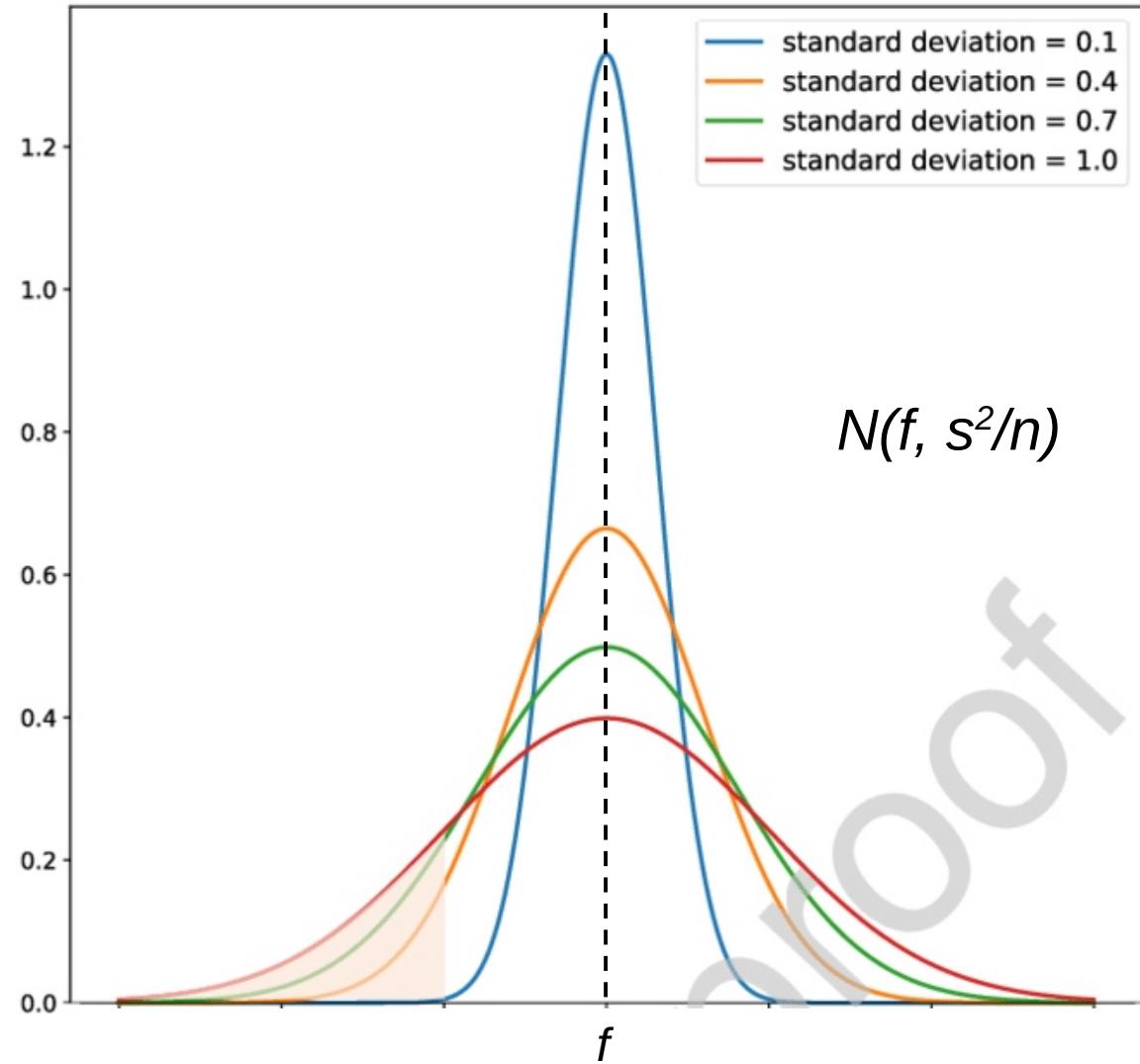
f -ed része Fideszre szavaz

95%-os konfidenciaszint





f -ed része Fideszre szavaz





Barnóczki Brigitta



Másolás



2082

A Republikon Intézet legfrissebb [kutatása szerint](#) januárban hibahatáron belül tovább nőtt a Tisza Párt előnye a Fidesz-KDNP-vel szemben, a teljes népesség körében 3 százalékponttal vezet Magyar Péter pártja, **26 százalékpontra** támogatottsága a Fidesz-KDNP 23-as eredményével szemben. Az MSZP-közeli intézet reprezentatív kutatása szerint a pártválasztók körében 37 százalék szavazna a Tiszára, 32 pedig a Fidesz-KDNP-re.

Módszertan: A kutatás 1000 fő telefonos megkérdezésével készült 2025. január 14–22. között. A kutatás nem, életkor, iskolai végzettség és településtípus szerint reprezentatív az ország felnőtt lakosságára. A hibahatár: +/- 3,2%.

Az intézet oldalán is elérhető [kutatás](#) szerint a vizsgált időszakban 39-32 volt az állás a biztos szavazó pártválasztók körében, ami az intézet szerint azt jelenti, hogy a Tisza magabiztosan vezet, de a növekedési üteme lassult. Az intézet azt

telex

Kövess minket Faceboo

Követem!

Legfontosabb



KÜLFÖLD

Mire a tűzoltó a negyedik és emeletek már emberek a fel színteken sik



A modern kor nyolcvan éve

1. gyakorlat: partok.csv elemzése

- <https://colab.research.google.com/>
- <https://ruzsaz.github.io/partok.csv>
- Határozzuk meg a két párt támogatottságát.
- Számoljuk ki a 95%-os, illetve 99%-os konfidenciaintervallumokat.
- Hogyan változik mindez, ha csak 100 fős mintát használunk?



Untitled2.ipynb ☆

Fájl Szerkesztés Nézet Beszúrás Futtatókörnyezet Eszközök Súgó

+ Kód + Szöveg ✓ RAM Lemez ▾

Létrehozás load data from https://ruzsaz.github.io/partok.csv

< 1 of 1 > Körültekintően használja a kódot

[1] # prompt: load data from https://ruzsaz.github.io/partok.csv

```
library(readr)
df <- read_csv("https://ruzsaz.github.io/partok.csv")
head(df)
```

Rejtett kimenet megjelenítése

Létrehozás calculate the proportion of preferencia="fidesz" within the df

< 1 of 1 > Módosítások visszavonása Körültekintően használja a kódot

prompt: calculate the proportion of preferencia="fidesz" within the df

```
nrow(df[df$preferencia == "fidesz", ])/nrow(df)
```

0.270036991368681

Létrehozás calculate the 95% confidence interval

< 1 of 1 > Módosítások visszavonása Körültekintően használja a kódot

[12] # prompt: calculate the 95% confidence interval

```
# Calculate the 95% confidence interval
conf_interval <- prop.test(nrow(df[df$preferencia == "fidesz", ]), nrow(df), conf.level = 0.95)

# Print the confidence interval
print(conf_interval)
```

1-sample proportions test with continuity correction

```
data: nrow(df[df$preferencia == "fidesz", ]) out of nrow(df), null probability 0.5
X-squared = 170.63, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.2400260 0.3022599
```

HF 1

Három közvélemény-kutató megmérte egy párt támogatottságát, az alábbi eredményekkel:

- 1) 36.2%, 600 fős minta
- 2) 29.8%, 2500 fős minta
- 3) 31.4%, 1000 fős minta

Ezek hibahatáron belüli eltérések?

Az eredmények összesítésével milyen eredményt kapunk? Milyen hibahatárral?





2. gyakorlat: horoszkop.csv elemzése

- <https://ruzsaz.github.io/horoszkop.csv>
- 2010.01.01 – 2022.04.30 között elhunyt, magyar taj számmal rendelkezők adatai.
- Nézzük meg az adatok struktúráját.
- Számoljuk ki a várható élettartamot a saját születési hónapunkban születettekre. Mekkora a 95%-os konfidenciaintervallum?
- HF2: alakítsuk függvényé az előzőt, számoljuk ki az összes hónapra, ábrázoljuk az eredményt.



 + Kód  + Szöveg RAM  Lemez Bezárás Létrehozás load data from <https://ruzsaz.github.io/horoszkop.csv> 1 of 1  Módosítások visszavonása [Körültekintően használja a kódot](#)[13] # prompt: load data from <https://ruzsaz.github.io/horoszkop.csv>

```
library(readr)
df <- read_csv('https://ruzsaz.github.io/horoszkop.csv')
head(df)
```

 Rejtett kimenet megjelenítése Létrehozás select rows where the value of szul_ho = 8, and store in aug 1 of 1    Körültekintően használja a kódot0  # prompt: select rows where the value of szul_ho = 8, and store in aug

```
mp
aug <- df |> dplyr::filter(szul_ho == 8)
head(aug)
```

 Rejtett kimenet megjelenítése Létrehozás calculate the mean of kor from the data frame aug, and caculate the 95% confidence interval 1 of 1  Módosítások visszavonása [Körültekintően használja a kódot](#)0  [23] # prompt: calculate the mean of kor from the data frame aug, and caculate the 95% confidence interval

```
mp
library(dplyr)
# Calculate the mean of 'kor' for August birthdays
mean(aug$kor, na.rm = TRUE)

# Calculate the 95% confidence interval
t.test(aug$kor, conf.level = 0.95)$conf.int
```

 27229.5188295536
27203.3899539545 · 27255.6477051526



Eldobott sorszámok:
9, 15, 16, 121, 248

Hány vásárló volt aznap?



German Tank Problem



3. gyakorlat: 9, 15, 16, 121, 248

- Mi a matematikai modell?
- https://en.wikipedia.org/wiki/German_tank_problem
- Ha m a maximális érték, k az elemszám, akkor a várható hibát minimalizáló torzítatlan becslés:

$$m(1+1/k) - 1$$

95%-os konfidencia-intervallum: $[m, m/0.05^{1/k}]$

- Írunk függvényt R-ben, ami ezeket kiszámolja!



+ Kód + Szöveg

RAM Lemez

Létrehozás

Solve the german tank problem from the numbers 9, 15, 16, 121, 248. Calculate the 95% confidence interval.



Bezáras

< 1 of 1 >

Módosítások visszavonása Körültekintően használja a kódot

0 mp

```
# prompt: Solve the german tank problem from the numbers 9, 15, 16, 121, 248. Calculate the 95% confidence interval.

# Function to calculate the German Tank Problem solution
german_tank_problem <- function(observations) {
  m <- max(observations)
  n <- length(observations)
  estimate <- m + (m / n) - 1

  # Calculate the 95% confidence interval
  k <- m
  lower_bound <- m
  upper_bound <- m / ((0.05)^((1/n)))

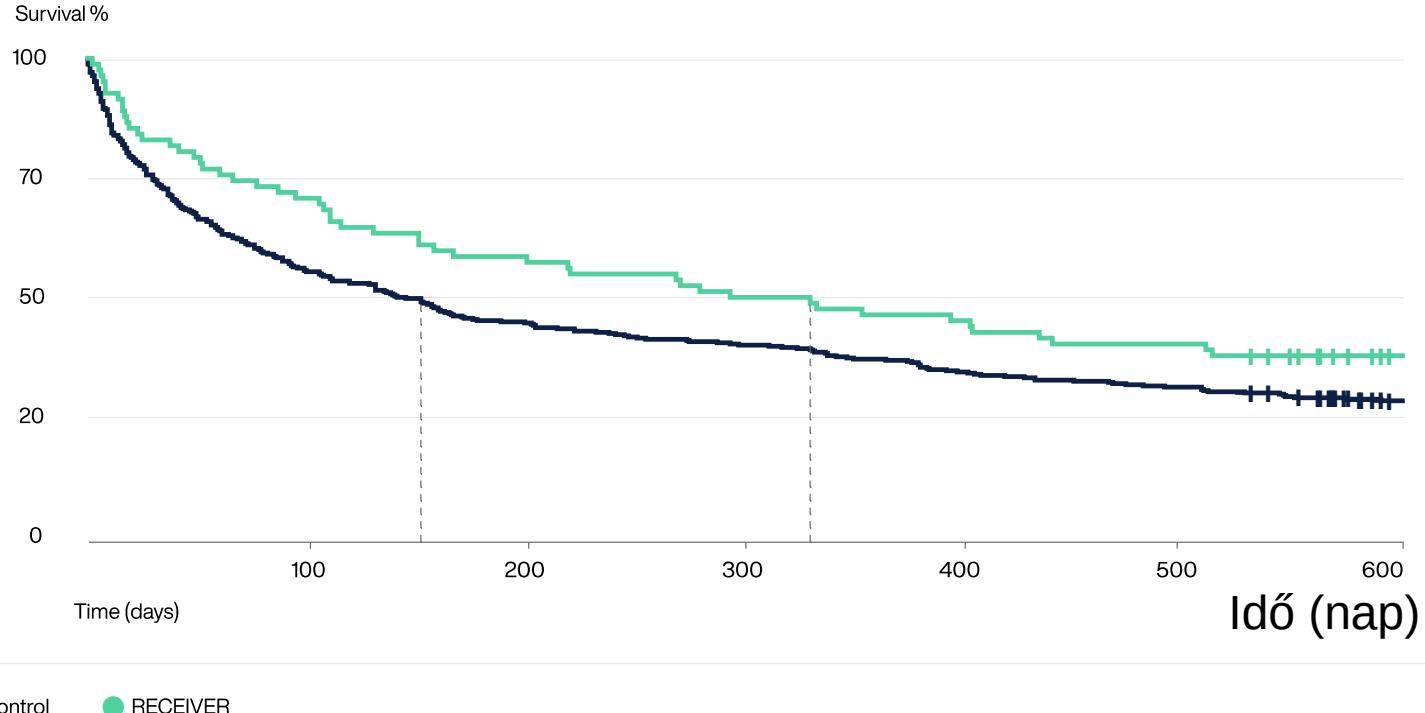
  list(estimate = estimate, lower_bound = lower_bound, upper_bound = upper_bound)
}

observations <- c(9, 15, 16, 121, 248)
result <- german_tank_problem(observations)

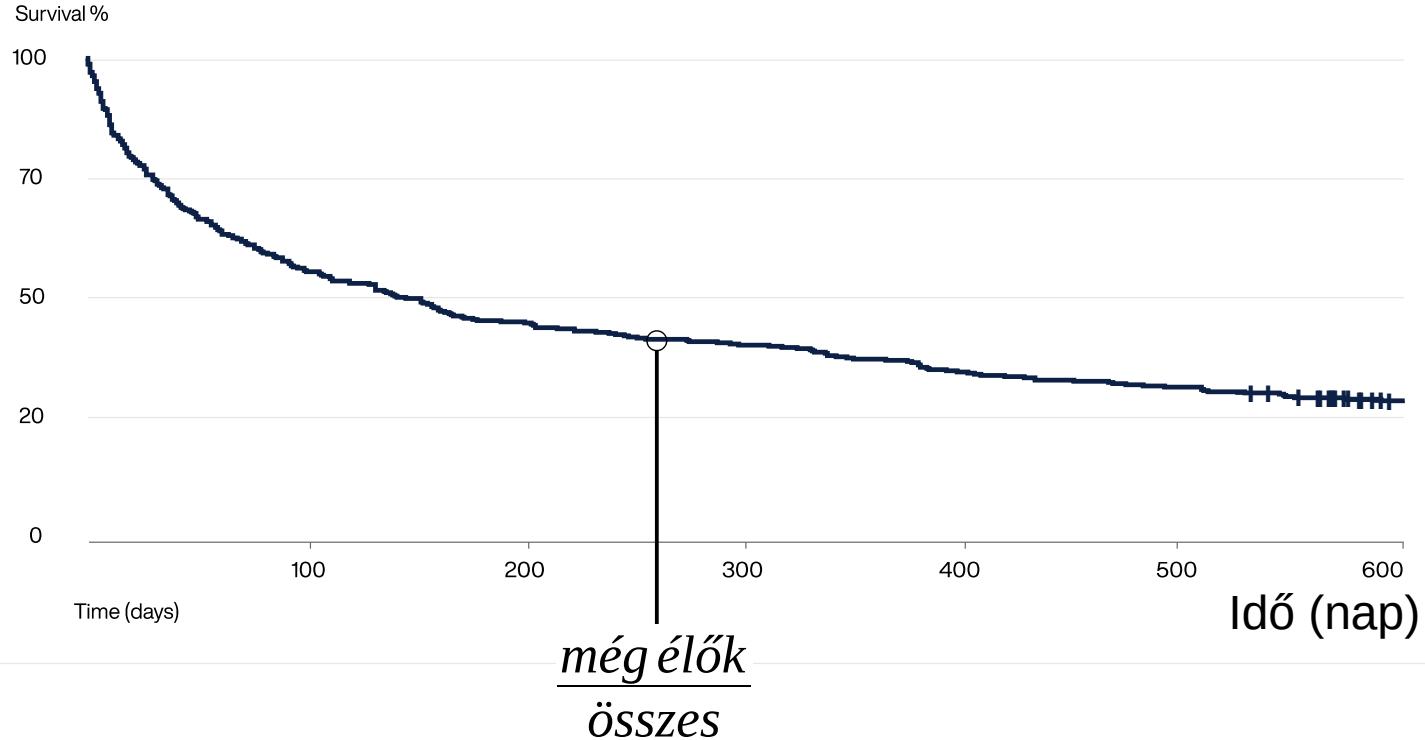
cat("Estimate:", result$estimate, "\n")
cat("95% Confidence Interval:", result$lower_bound, "-", result$upper_bound, "\n")
```

```
Estimate: 296.6
95% Confidence Interval: 248 - 451.4999
```

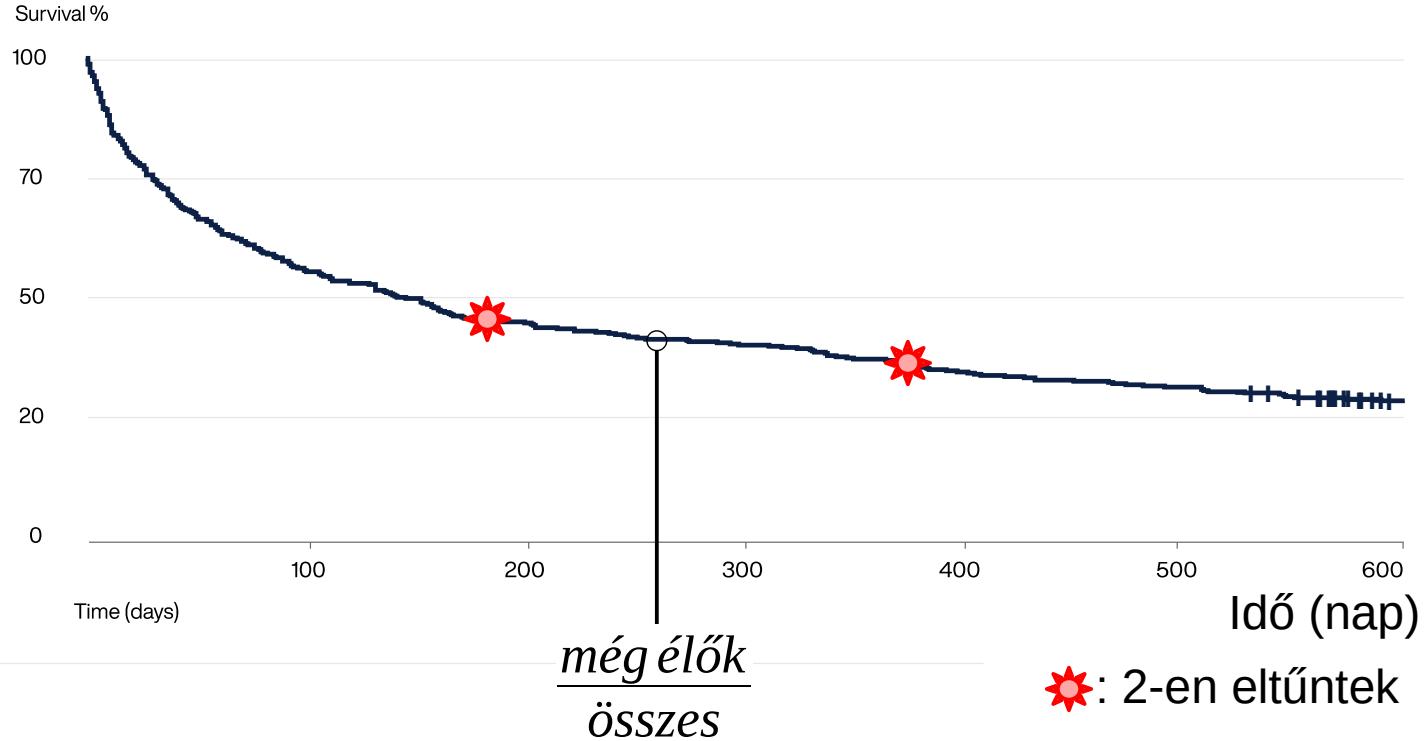
Kaplan-Meier túlélési görbe



Túlélési görbe (K.M. nélkül)



Túlélési görbe (K.M. nélkül)



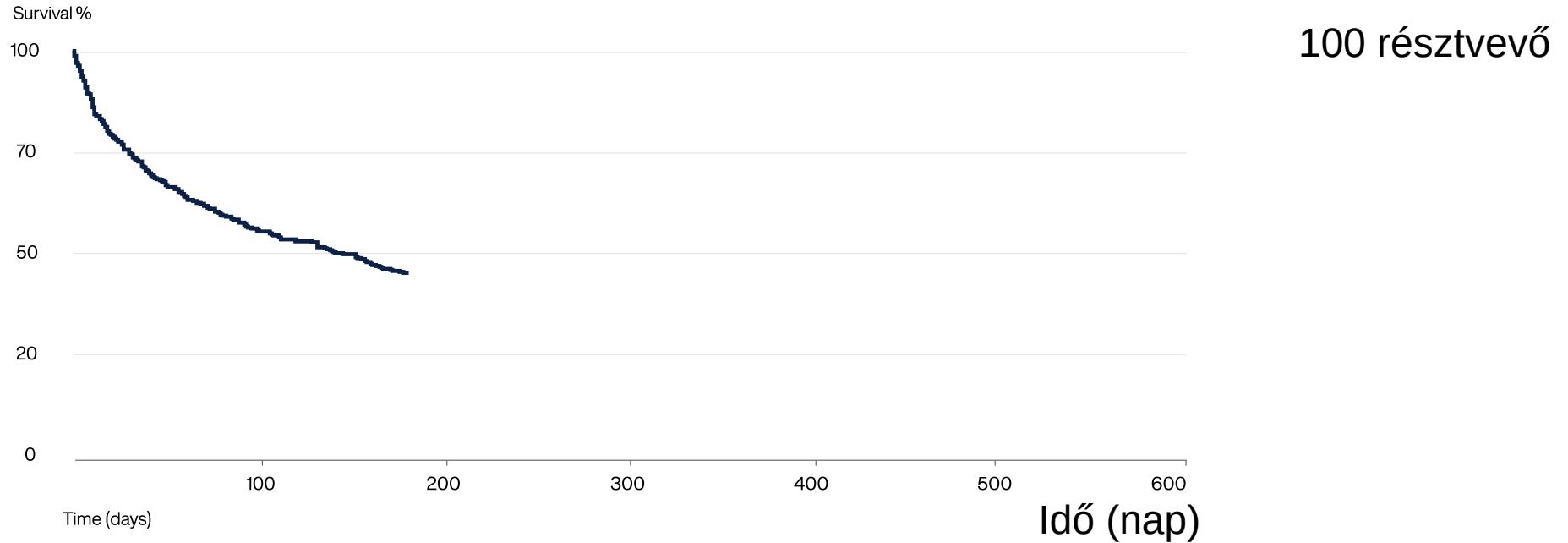
100 résztvevő

Idő (nap)

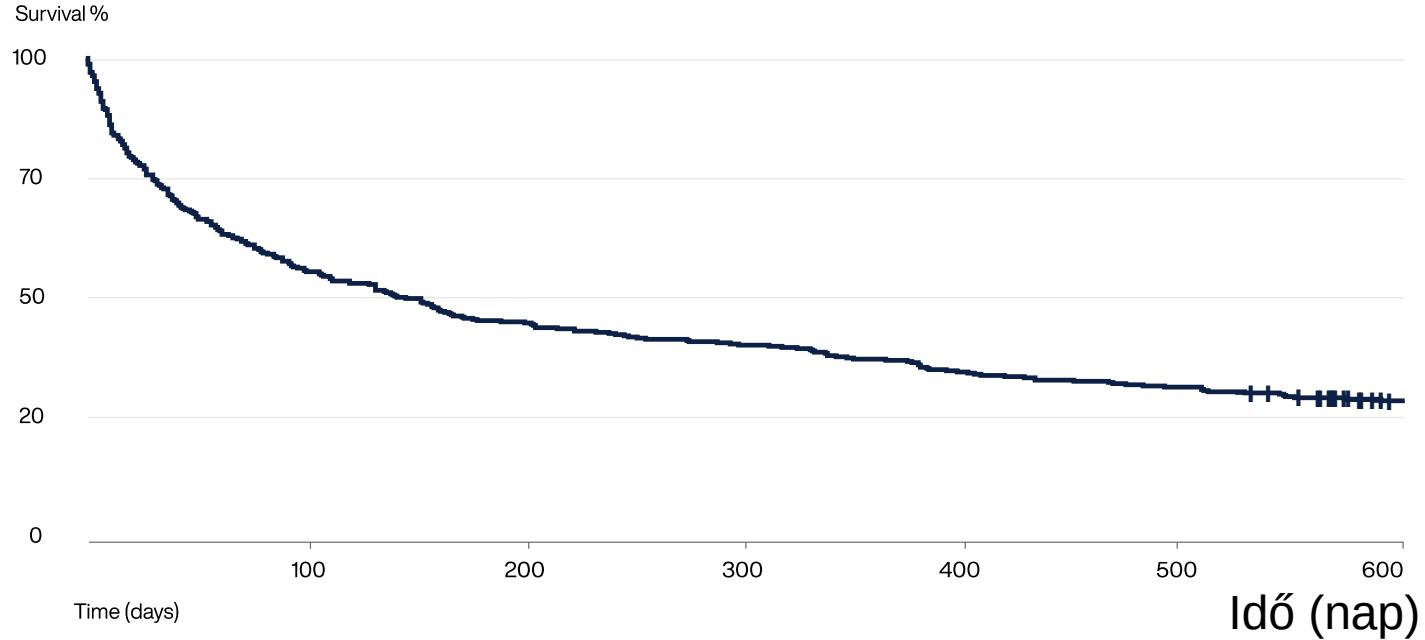
✳: 2-en eltűntek



1) Abbahagyjuk?



2) Letagadjuk a 2 résztvevőt?

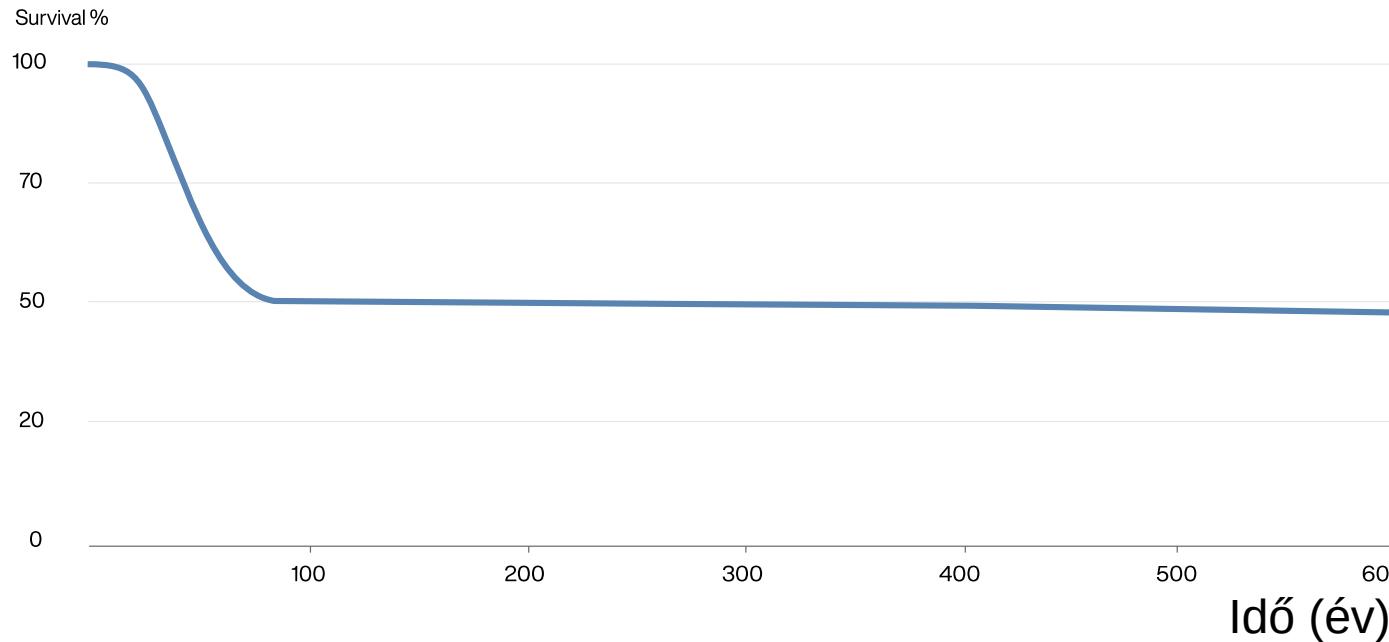


~~100 résztvevő~~

98 résztvevő



Örök élet!!!*

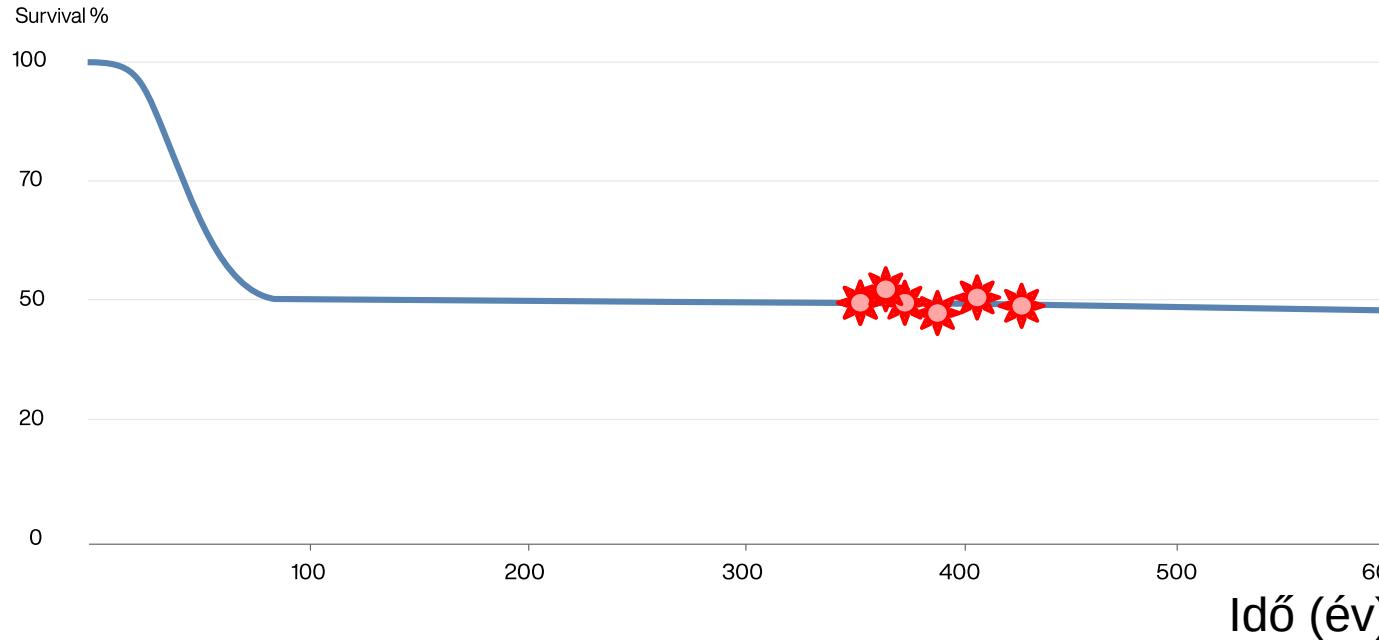


1000 résztvevő



*(50%-os hatékonysággal)

Örök élet!!!*



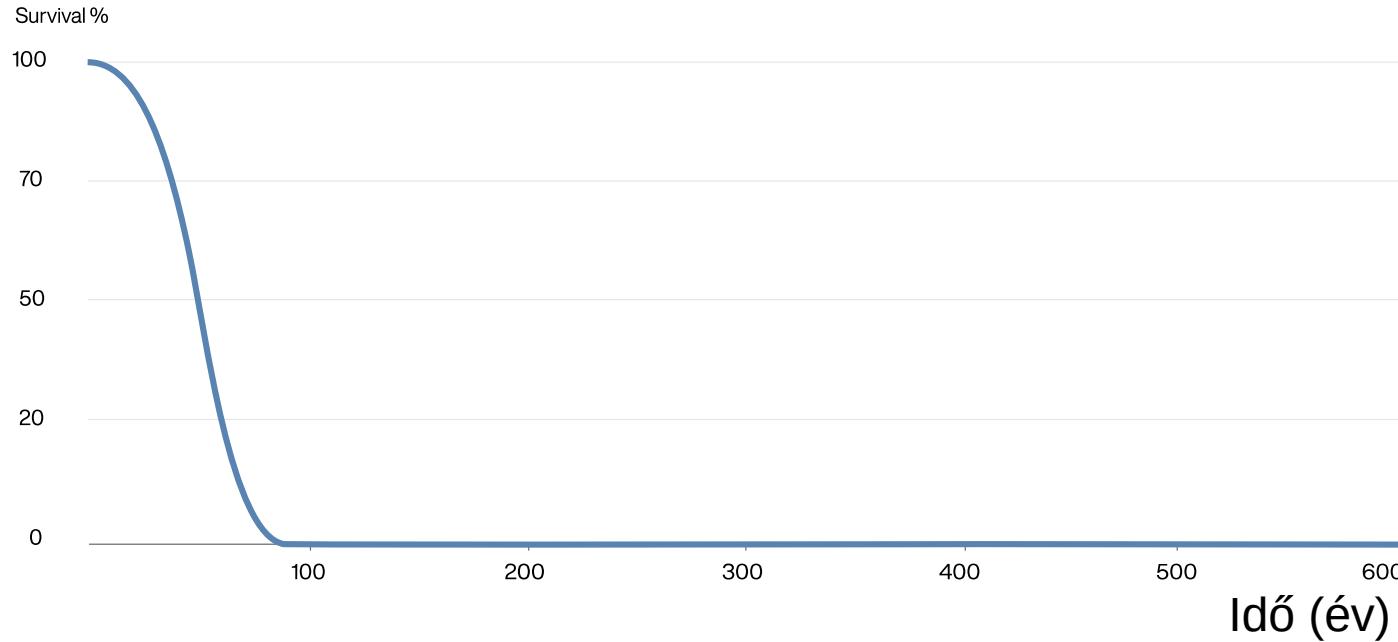
1000 résztvevő

✳️: 500-an eltűntek



*(50%-os hatékonysággal)

Örök élet!!!*

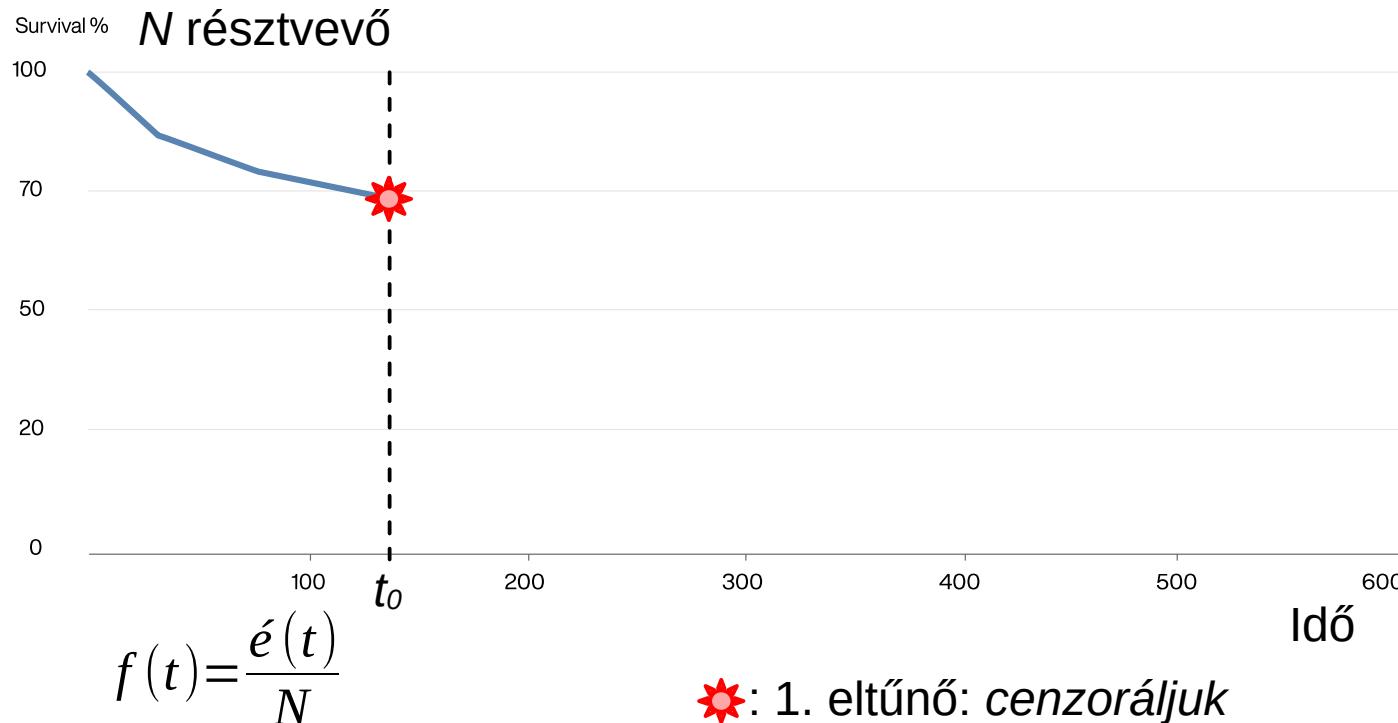


~~1000 résztvevő~~
500 résztvevő

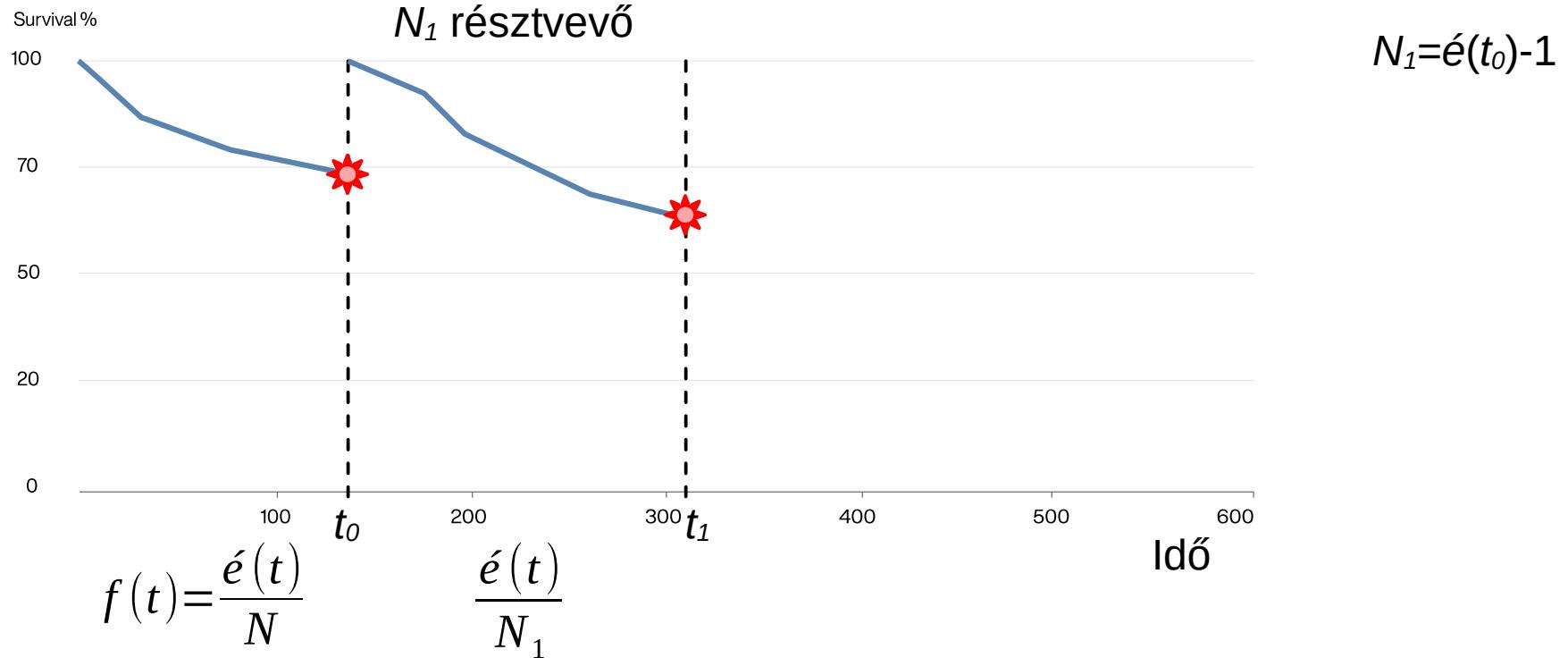


~~*(50%-os hatékonysággal)~~

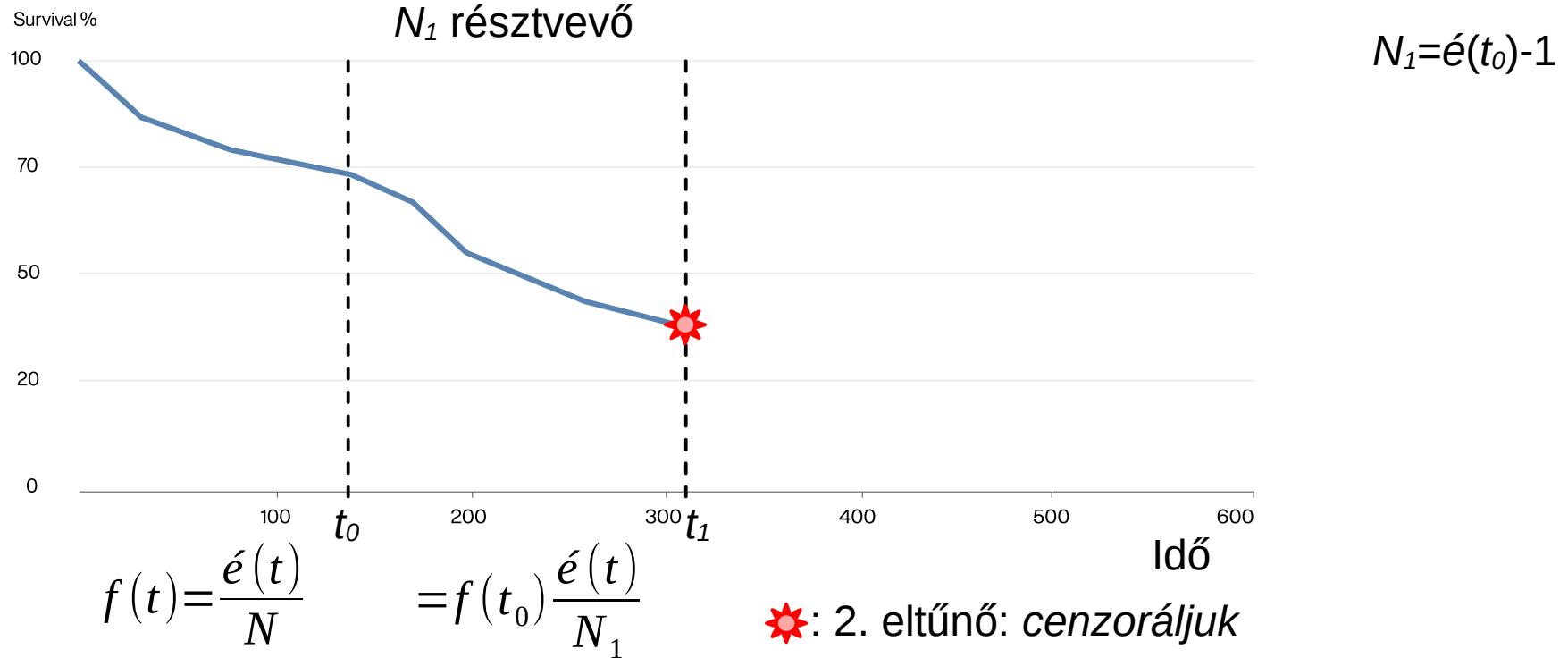
Kaplan-Meier



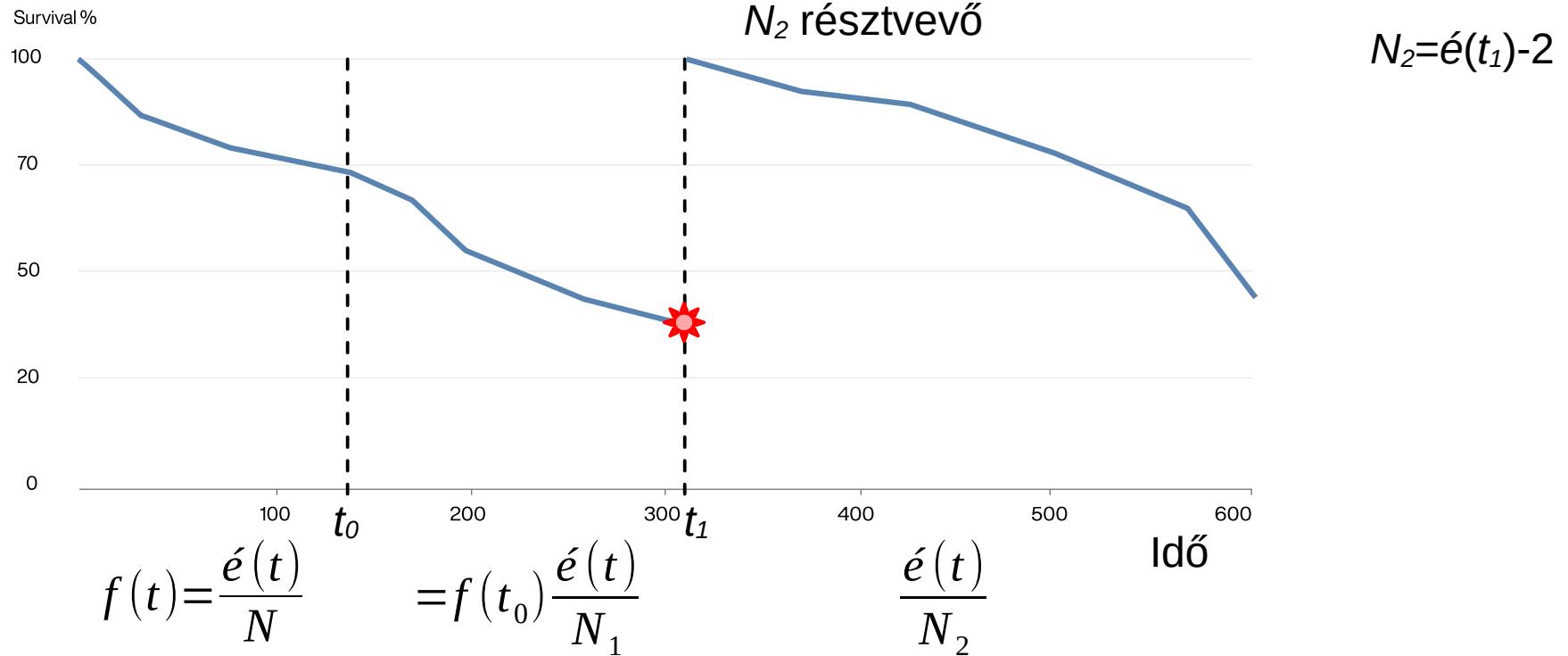
Kaplan-Meier



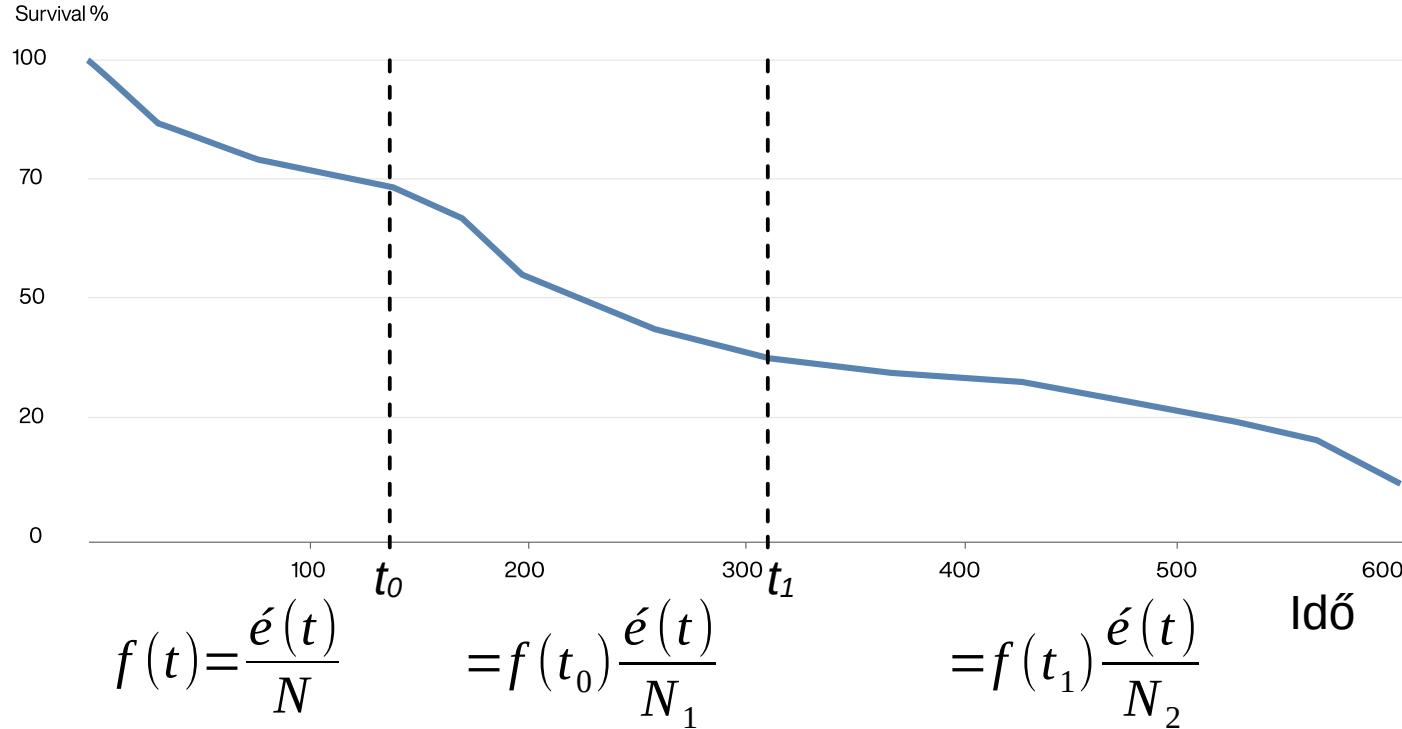
Kaplan-Meier



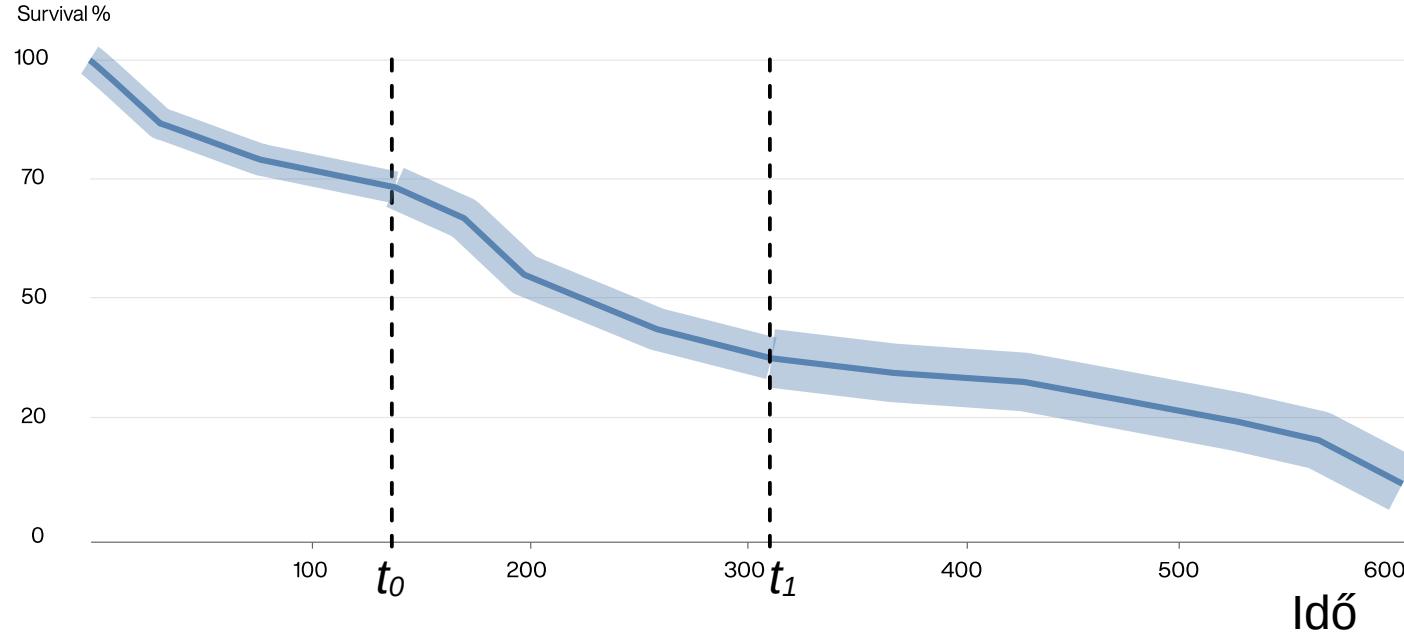
Kaplan-Meier



Kaplan-Meier



Kaplan-Meier becslés



4. gyakorlat: tuleles.csv elemzése

- <https://ruzsaz.github.io/tuleles.csv>
- 3 oszlop:
 - id: azonosító,
 - ido: esemény ideje (év),
 - esemeny: 1 – meghalt, 0 – eltűnt.
- Rajzoljuk fel a Kaplan-Meier túlélési görbét.
- Nézzük meg, hogy nézne ki a görbe, ha a cenzorálandó eseményeket inkább kidobnánk az alapadatokból.



+ Kód + Szöveg

✓ RAM Lemez

Létrehozás load file from <https://ruzsaz.github.io/tuleles.csv>



Bezáras

< 1 of 1 > Körültekintően használja a kódot

```
[1] # prompt: load file from https://ruzsaz.github.io/tuleles.csv

library(readr)
df <- read_csv("https://ruzsaz.github.io/tuleles.csv")
head(df)
```

Rejtett kimenet megjelenítése

Létrehozás make the kaplan.meier survival amalysis of df.



Bezáras

< 1 of 1 > Módosítások visszavonása Körültekintően használja a kódot

prompt: make the kaplan.meier survival amalysis of df.

```
library(survival)
library(ggplot2)

# Assuming 'time' and 'status' are the relevant columns in your dataframe
surv_object <- Surv(time = df$ido, event = df$esemény)

# Perform Kaplan-Meier analysis
km_fit <- survfit(surv_object ~ 1, data = df) # ~ 1 means no grouping variable

# Plot the survival curve
autoplot(km_fit, xlab = "Time", ylab = "Survival Probability", main = "Kaplan-Meier Survival Curve")
```

Rejtett kimenet megjelenítése

Létrehozás make the y axis 0-100% in the previous graph, and save it in a file as a 2000px wide



Bezáras

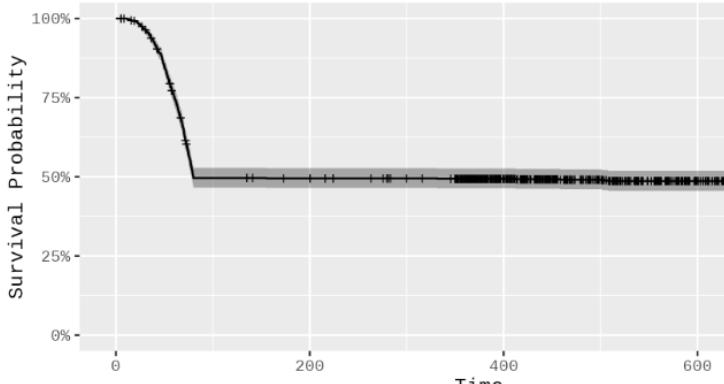
< 1 of 1 > Körültekintően használja a kódot

```
[39] # prompt: make the y axis 0-100% in the previous graph, and save it in a file as a 2000px wide and 1000px tall png

ggsave("survival_curve.png", width = 2000, height = 1000, units = "px",
       plot = autoplot(km_fit, xlab = "Time", ylab = "Survival Probability", main = "Kaplan-Meier Survival Curve") +
         scale_y_continuous(limits = c(0, 1), labels = scales::percent))
```

Gemini survival_curve.png

Kaplan-Meier Survival Curve



<https://ruzsaz.github.io/stat1.pdf>

HF1: Három közvélemény-kutató megmérte egy párt támogatottságát, az alábbi eredményekkel:

- 1) 36.2%, 600 fős minta
- 2) 29.8%, 2500 fős minta
- 3) 31.4%, 1000 fős minta

Ezek hibahatáron belüli eltérések?

Az eredmények összesítésével milyen eredményt kapunk? Milyen hibahatárral?

HF2: Hajtsuk végre a <https://ruzsaz.github.io/horoszkop.csv> elemzését R-ben, az alábbi lépésekkel:

- 1) Írunk függvényt, amely egy bemenő paraméter (1-12-ig) alapján kiszámolja az adott hónapban születettek élettartamát, illetve a 95%-os konfidenciaintervallumot.
- 2) Csinálunk egy 12 elemű vektort, ami 1-12-ig tartalmazza a számokat.
- 3) Hívjuk meg a függvényt a vektor elemeire.
- 4) Ábrázoljuk az eredményeket (becslést és konfidenciaintervallumot).

