Mario Kart 8 Statisztikák

Többváltozós adatelemzési modellek (MAME039LMSB) házidolgozat

Fazekas Márk Máté

EUV0J3

**Tartalomjegyzék**

[Bevezetés 2](#_Toc125753496)

[Az adatbázis bemutatása 3](#_Toc125753497)

[Score elemzése (Leíró statisztikai elemzés) 5](#_Toc125753498)

[Size (nem numerikus) 5](#_Toc125753499)

[Intervallumbecslés és Hipotézisvizsgálat 6](#_Toc125753500)

[Kétváltozós kapcsolatvizsgálat 7](#_Toc125753501)

# Bevezetés

A Mario Kart egy Nintendo által fejlesztett és forgalmazott versenyzős játék.A Mario Kart 8 eredeti kiadását 2014-ben hozták forgalomba, a Delux változat (ami az előző frissítése), 2017-ben került a nyilvánosság elé. A mai napig frissíti a Nintendo, ezzel a felhasználókat visszacsábítva időről időre. Ez a játék nem csak az átlagos felhasználók, de az úgynevezett „speed-runnerek” által is közkedvelt. Ezek olyan felhasználók, akik minél jobb időt szeretnének elérni, a bizonyos versenypályákon.

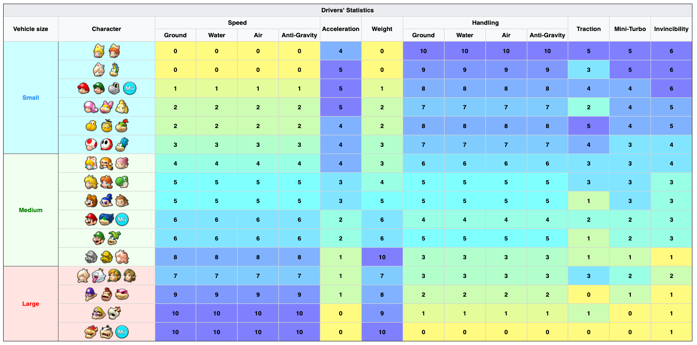
Egy versenyző eredményét a tradicionális futamokon több dolog is befolyásolhatja, de a versenyek a minél jobb időkért általában az időmérő futamokon szokott történni. Itt tényleg csak a versenyzőn múlik minden. Az egyik fontos döntés, amit a versenyző meghozhat, hogy milyen „konfigurációval” fog versenyezni.

Minden versenyzőnek választania kell egy sofőrt (driver), egy járművet (kart), egy fajta kereket (tire) és egy ernyőt (glider). Ennek a 4 komponens összetételével különböző járművek kreálhatóak, amik a különböző tulajdonságaik miatt, különböző eredményekhez vezethetnek. A játék ezeket a tulajdonságokat számokkal jellemzi, és grafikonokon vizualizálja a játékos számára. A továbbiakban azt fogom vizsgálni, hogy átlagos játékosoknak, milyen kombináció lehet a leghatékonyabb. Mivel több felületen folyik a verseny a játékban (szárazföld, víz, levegő, mesterséges gravitáció), ezért a pontos elemzéshez a pályák pontos összetétele is szükséges lenne, ami jelenleg nem elérhető, így az egyszerűség kedvéért, csak a szárazföldi paramétereket elemzem.

# Az adatbázis bemutatása

Forrás: <https://www.kaggle.com/datasets/marlowspringmeier/mario-kart-8-deluxe-ingame-statistics> (Letöltés dátuma: 2023.01.21).

A kaggle-n elérhető adatbázissal könnyebb volt elkezdeni dolgozni, de a MarioWikin, gyakrabban frissített adatok érhetőek el.



ábra 1 Driver statisztikák MarioWikin

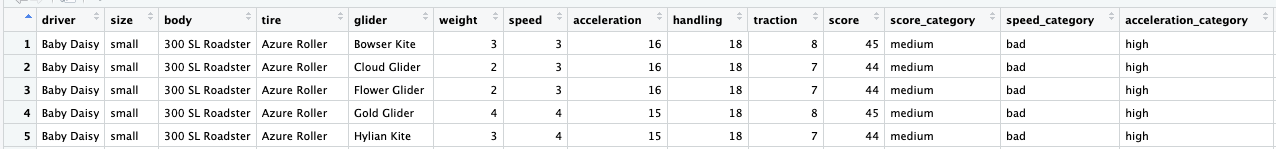
Az eredeti adatbázis négy táblából áll (driver, kart, tire, glider). A négy táblának Descartes-szorzata és oszlopainak egyszerűsítésével generáltam az elemzett adatbázist.

Egy megfigyelési egység egy lehetséges kombináció, amit a játékos választhat (43\*40\*21\*14 = 505 680 megfigyelés):

* Driver
  + A választott karakter (pl.: Mario)
  + 43 különböző érték
* Kart
  + A választott jármű (pl.: Standard Kart)
  + 40 különböző érték
* Tire
  + A választott kerék (pl.: Standard)
  + 21 különböző érték
* Glider
  + A választott ernyő (pl.: Super Glider)
  + 14 különböző érték

És a hozzá tartozó mutatók:

* Size
  + A driver mérete
  + Ordinális minőségi változó
    - Lehetséges értékek: small / medium / large
* Weight
  + A konfiguráció súlya. A nehezebb konfigurációkkal, ha kiütünk egy játékost, tovább tart az ellenfélnek újra indulnia. A nehezebb játékosokat mikor kiütik, ők is nehezebben indulnak újra.
  + Különbségi (intervallum) mérési skála
* Speed
  + A konfiguráció maximum sebessége.
  + Különbségi (intervallum) mérési skála
* Acceleration
  + A konfiguráció gyorsulása. A maximum sebesség növekedés képkockánként
  + Különbségi (intervallum) mérési skála
* Handling
  + A konfiguráció sebessége kanyarodás közben.
  + Különbségi (intervallum) mérési skála
* Traction
  + A konfiguráció irányíthatósága
  + Különbségi (intervallum) mérési skála
* Score
  + Speed + Acceleration + Handling + Traction
  + Különbségi (intervallum) mérési skála
* Score\_Category
  + 5 kategóriába sorolt összepont szöveges megjelenítése
* Speed\_Category
  + 5 kategóriába sorolt speed szöveges megjelenítése
* Acceleration\_Category
  + 5 kategóriába sorolt acceleration szöveges megjelenítése



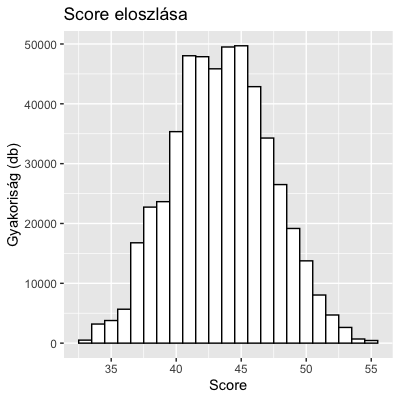
ábra 2 df

# Score elemzése (Leíró statisztikai elemzés)

A score 33 és 55 közötti értékek lehetnek. Az elemszám itt is 505 680, mivel minden konfigurációnak van pontja.

A hisztogramról és a gyakorisági ábráról látható, hogy a leggyakoribb érték a 45 (módusz).

A medián 43 (azaz a konfigurációk fele ennél rosszabb és a másik fele ennél jobb) és az átlag 43.48. Azaz, egy átlagos konfiguráció 43.48 pontos.



ábra 3 hisztogram

A hisztogramról és az átlag, módusz és a medián viszonyából feltételezem, hogy az adat szimmetrikus eloszlású. Az α3 (skew) értéke 0.04, ez igazolja a feltételezést, hogy szimmetrikus eloszlású. Ez ugye azt is jelenti, hogy a változó értékei véletlen hatások összegződésével állnak elő.

Az α4 (kurtosis) értéke -0.32, ami azt jelenti, hogy a változó eloszlása a normális eloszlásnál lapultabb.

A szórás 3.8, azaz egy véletlenszerűen kiválasztott konfiguráció értéke 3.8 ponttal fog eltérni az átlagtól. (Nem beépített R sd, mivel ismerjük a teljes populációt)

A konfigurációk negyede több, mint 46 pontos, míg a negyede kevesebb, mint 41 pontos.

Chart, histogram

Description automatically generated

ábra 4 dobozábra

Felső kerítés 53.5 (46+1.5\*5). Ez azt jelenti, hogy az ennél nagyobb score értékek kilógóan magasak. A gyakorisági táblából látható 1134 ilyen rekord van (ez az elemek 0.2%-a). Az alsó kerítés 33.5 (41-1.5\*5). Ebben a kategóriában 3696 rekord van (az elemek 0.7%-a).

## Size (nem numerikus)

Az értékkészlete 3 különböző érték. A módusz és a medián is „medium”. A medián értelmezhető, mert a három érték egyértelműen sorba rendezhető (ordinális). Ez azt jelenti, hogy konfigurációk leggyakrabban közepes méretű sofőröket tartalmaz, illetve, hogy a sorbarendezett elemek középső értéke is „medium”.

# Intervallumbecslés és Hipotézisvizsgálat

Csináltam egy 48 elemű mintát, amivel szimulálom egy „háziverseny” összetételét.

Text

Description automatically generated with medium confidence

ábra 5 Intervallumbecslés

A kis méretű sofőrökkel rendelkező konfigurációk 99%-os valószínűséggel legalább 38,8 pontosak lesznek és legfeljebb 44,1. A közepes méretű sofőrökkel rendelkező konfigurációk 99%-os valószínűséggel legalább 41,2, és legfeljebb 45,2 pontosak lesznek. A nagy sofőrök konfiguráció szintén 99%-os valószínűséggel legalább 38,9 és legfeljebb 48,0 lesz. Azt látjuk, hogy a score nem különbözik szignifikánsan, hiszen bármelyik csoport intervalluma fedi a többit.

A vizsgált mintában a score átlaga 42,81.

Feltételezem, hogy a teljes sokaságban az átlag legalább 45 (null hipotézis). Alternatív hipotézisem, hogy az átlag kevesebb, mint 45. (baloldali próba).

Text

Description automatically generated

ábra 6 t-test

Mivel a kapott p-érték (0.0001794) kevesebb, mint a legkisebb szokásos szignifikancia szint (1%), H0-t elutasítom, és az alternatív hipotézisemet elfogadom.

# Kétváltozós kapcsolatvizsgálat

## Vegyes kapcsolat

Vizsgáljuk a sofőr méretét (size) és a score közötti összefüggést. A kapcsolat típusa vegyes (minőségi-mennyiségi), ezért egy csoportosított doboz ábrán ábrázoljuk. Láthatjuk, hogy a súlyosabb versenyzők, általában magasabb értékű konfigurációkhoz társulnak, míg a könnyű sofőrök átlagosan kevesebb pontos konfigurációt érnek el. Mivel a középső súlycsoportnál a legkisebb score nagyobb, mint a másik kettő csoportnál, ezért mondhatjuk, hogy aki biztosan nem választaná a legrosszabb konfigurációt, az kezdje egy középsúlyú sofőrrel.

Chart, box and whisker chart

Description automatically generated

ábra 7 kétváltozós boxplot

SSB: 233852 (size átlag távolsága a score átlagtól). SSR: 7102769 (az adott méret score értékének távolsága a saját size csoportjuk átlagától). SST: 7336621 (a konfigurációk score-jának távolsága a score főátlagtól).

Text

Description automatically generated

ábra 8 aov

Variancia-hányados kb 3.19%: a méret a konfiguráció pontjának alakulásának (variancájának) 3,19%-át magyarázza a megfigyelt mintában. Ez egy gyenge kapcsolat, mivel a variancia hányados 10%-nál kevesebb. Szórás hányados 0,18, szóval a kapcsolat továbbra is gyenge marad.

Chart

Description automatically generated with medium confidence

ábra 9 F-próba

Mivel a nominális változó mindhárom csoportjában áll a nagy minta feltétel, így továbbmegyünk az F-próba p-értékének kiszámításához. Az eredmény alapján a p-érték kisebb, mint 2×10−16. Ez kisebb még a legkisebb szokásos szignifikancia-szintnél, az α=1%-nél is, így egyértelműen és stabilan elfogadható a H1, ami szerint a méret magyarázóereje a végső pontokra nézve szignifikánsan több a sokaságban is, mint 0. Azaz a magyarázóerő nem a mintavételi hiba műve.

Végkövetkeztetésül azt mondhatjuk el, hogy a Mario Kart 8 konfigurációjában nincsen kapcsolat a sofőr mérete és a konfiguráció végső pontja között.

## Asszociációs kapcsolat

Vizsgáljuk meg a size és a speed\_category kapcsolatát. A kapcsolat típusa asszociációs (minőségi-minőségi), ezért egy halmozott oszlop diagramon ábrázoljuk. A táblából látszik, hogy az „excellent” speed\_categoryban 90,5%-ban „large” méretű sofőrök találhatóak.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generatedChart

Description automatically generated

Cramer-együttható 0,57 – a minőségi változók között közepes kapcsolat áll fent.

A Khi-négyzet vizsgálat eredménye alapján a p-érték kisebb, mint 2×10−16. Ez kisebb még a legkisebb szokásos szignifikancia-szintnél, az α=1%-nál is, ezek szerint egyértelműen és stabilan elfogadható lenne a H1, ami szerint a size magyarázóereje a speed\_category-ra nézve szignifikánsan több a sokaságban is, mint 0, viszont a Khi-négyzet próba előfeltétele (mi szerint legalább 5 elemnek kell lennie minden kategóriában) nem teljesült.

Így összességében kijelenthetjük, hogy a méret és a sebesség kategóriában fennálló közepes kapcsolat áll fent, viszont ezt nem általánosíthatjuk a nem vizsgált egyedekre.

## Korrelációs kapcsolat

Vizsgáljuk meg a sebesség és a gyorsulás közötti kapcsolatot. A kapcsolat típusa korrelációs (mennyiségi-mennyiségi), ezért pontdiagramon ábrázoljuk.

A picture containing scatter chart

Description automatically generated

ábra 10 Speed vs Acceleration

Az ábrán is látható, hogy ellentétes kapcsolat áll fent a két változó között. A -0,84-es **korreláció** negatív előjele is erről árulkodik. Szóval, ha növelni szeretnénk a sebességet, akkor várhatóan csökken a gyorsulásunk. Mivel az abszolút érték 0,7-nél nagyobb, ezért erős/szoros kapcsolatot ír le a korreláció. Tehát azt mondhatjuk, hogy a sebesség és gyorsulás közti kapcsolat ellentétes irányú és erős. A sebesség kb. 71%-ban magyarázza a gyorsulást.

A regressziós egyenes egyenlete

Azaz a tengelymetszete 20,86 és a meredeksége -0.97, azaz a sebesség növelése eggyel várhatóan -0,97 gyorsulás csökkenéssel jár.

p-értékünk <2e-16, szinte nulla, ami kisebb, mint a legkisebb szokásos szignifikancia szint. Szóval minden szokásos szignifikancia-szinten elutasítható az a H0, miszerint a regressziós egyenes meredeksége a megfigyelt adatokon túli világban 0 lenne. A regressziónk új megfigyeléseken, új konfigurációkon is használható.

A reziduális standard hiba 1.703. azaz egy regressziós becslés a sebesség növelésével várható gyorsulás csökkenése várhatóan ±1,703-mal tér el a valós gyorsulás mértékétől.

Chart

Description automatically generated

A kutatás folyamán ezt a relációt találtam a legérdekesebbnek, mert a legtöbb profi ezen a pontdiagram alapján választ konfigurációt, mivel ez a kettő a legfontosabb statisztika és talán a legkönnyebben értelmezhető is. Itt gyakran emlegetik a Pareto-hatékonyságot, hiszen érdemes úgy választani, hogy ne legyen azonos sebességnél gyorsabban gyorsuló, vagy azonos gyorsuláson magasabb végsebességgel rendelkező konfiguráció. Pl. egy 20-as gyorsulású, 0-s sebességű konfiguráció nem Pareto-hatékony, hiszen a 20-as gyorsulással elérhető akár 2-es sebességű konfiguráció is. A játékban a profik ilyen Pareto-hatékony konfigurációkat használnak, a játékstílusuknak és a pályáknak megfelelően választva (ahol sok az egyenes rész, fontosabb a magasabb végsebesség, míg a kanyargós pályákon a gyorsulás fontosabb és eredményre vezetőbb).

Large:

Medium

Small

Ezek alapján a large csoportban a sebesség csökkentése eggyel 0.97-tel növeli a gyorsulást, míg a medium csoportban a sebesség csökkentése eggyel 1.12-vel növeli a gyorsulást.

A p-érték, mindig szinte nulla, ami kisebb, mint a legkisebb szokásos szignifikancia szint. Szóval minden szokásos szignifikancia-szinten elutasítható az a H0, miszerint a regressziós egyenes meredeksége a megfigyelt adatokon túli világban 0 lenne. A regressziónk új megfigyeléseken, új konfigurációkon is használható.

[ábra 1 Driver statisztikák MarioWikin 3](#_Toc125812773)

[ábra 2 df 4](#_Toc125812774)

[ábra 3 hisztogram 5](#_Toc125812775)

[ábra 4 dobozábra 6](#_Toc125812776)

[ábra 5 Intervallumbecslés 6](#_Toc125812777)

[ábra 6 t-test 7](#_Toc125812778)

[ábra 7 kétváltozós boxplot 7](#_Toc125812779)

[ábra 8 aov 8](#_Toc125812780)

[ábra 9 F-próba 8](#_Toc125812781)

[ábra 10 Speed vs Acceleration 9](#_Toc125812782)