Statisztikai Elemzés

Vida Péter

16 May 2018

Adatok

A adathalmaz erről a linkről származik (https://www.kaggle.com/abcsds/pokemon).

Az adataink a `data/Pokemon.csv` fájl tartalmazza. Soronként egy-egy pokémonról a következőket tudjuk:

- Name: a neve
- Type.1: az elsődleges típusa
- Type.2: a másodlagos típusa (lehet üres is)
- Total: az alábbi tulajdonágok összege:
  - HP: életerő pontok
  - Attack: támadóérték
  - Defense: védelmi érték
  - Sp..Atk: speciális támadás
  - Sp..Def: speciális védekezés
  - Speed: sebesség
- Generation: melyik generációból való.
- Legendary: legendás pokémon-e

```
data <- read.csv('data/Pokemon.csv')
labels(data)[2]
```

```
## [[1]]
##  [1] "X."          "Name"        "Type.1"      "Type.2"      "Total"
##  [6] "HP"          "Attack"      "Defense"     "Sp..Atk"     "Sp..Def"
## [11] "Speed"       "Generation" "Legendary"
```

Az adatok közül nem vettem figyelembe a pokémonok mega evolúcióját, mivel ez egy ideiglenes állapot. Ez 13 sor törlését jelentette. Hasonló elgondolásból a pokémonok primal formája is kikerült az adathalmazból. Illetve a Hoopa "Unbound" alakja is.

```
data <- data[!grepl("Mega",data$Name,fixed=TRUE),]
data <- data[!grepl("Primal",data$Name,fixed=TRUE),]
data <- data[!grepl("Unbound",data$Name,fixed=TRUE),]
```

Az adatokban a pokémonok, melyeknek különböző formái vannak, külön pokémonoknak számítanak. A duplikátumok elkerülése végett, ezeket a formákat kiszűrtem, a tulajdonságaikat pedig átlagoltam. Ilyen pokémonok voltak:

```
for (x in levels(factor(data$X.))) {
  rows <- data[data$X. == x,]
  if (nrow(rows) > 1) {
    data <- data[data$X. != x,]
    name <- rows$Name[1]
    name <- strsplit(gsub("(.)([[:upper:]])", "\\1 \\2", name),' ')[[1]][[1]]
    print(name)
    levels(data$Name) <- c(levels(data$Name),name)
    statlab <- c('Total','HP','Attack','Defense','Sp..Atk','Sp..Def','Speed')
    base <- rows[1,]
    base$Name <- name
```

```
    for (l in statlab) {
      base[[l]] <- mean(rows[[l]])
    }
    data[nrow(data)+1,] <- base
  }
}
```

```
## [1] "Deoxys"
## [1] "Wormadam"
## [1] "Rotom"
## [1] "Giratina"
## [1] "Shaymin"
## [1] "Darmanitan"
## [1] "Tornadus"
## [1] "Thundurus"
## [1] "Landorus"
## [1] "Kyurem"
## [1] "Keldeo"
## [1] "Meloetta"
## [1] "Meowstic"
## [1] "Aegislash"
## [1] "Pumpkaboo"
## [1] "Gourgeist"
```

A legendás pokémonok nagytöbbségében jobb tulajdonságokkal rendelkeznek, így az adathalmazunkat érdemes szétválasztani.

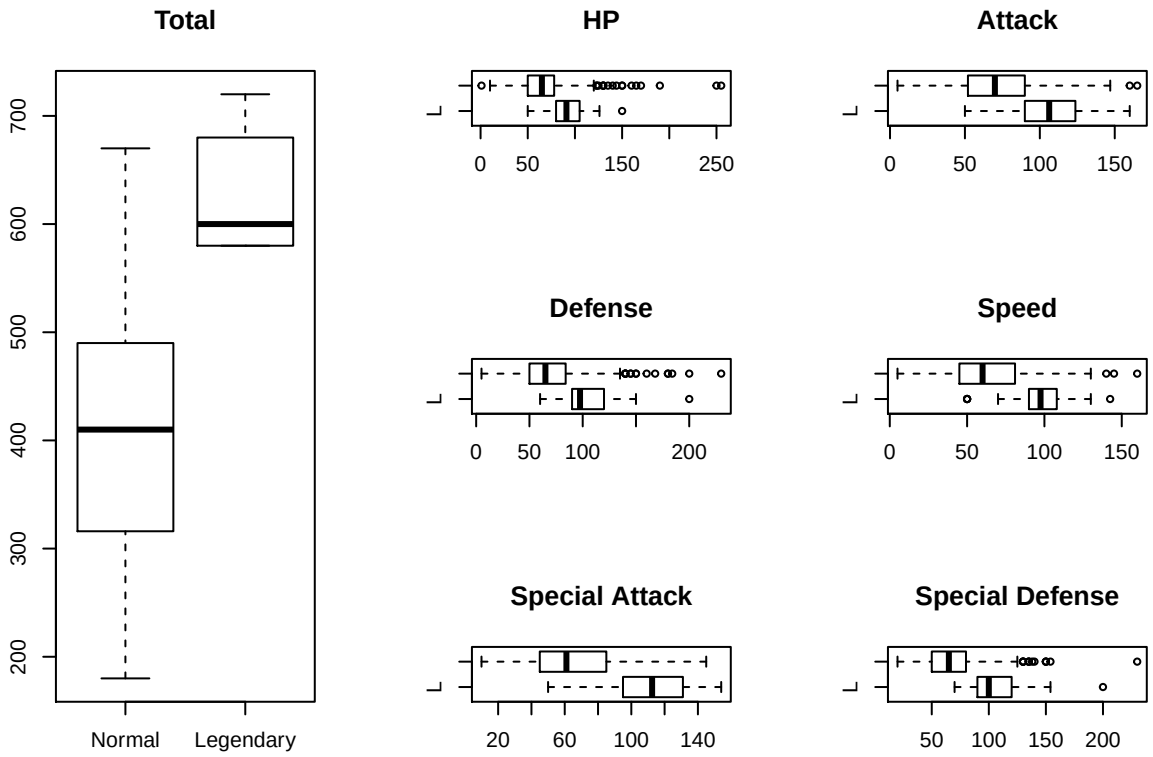
```
legendsplit <- split(data,data$Legendary)
normal <- legendsplit$False
legendary <- legendsplit$True
```

Így a normális és a legendás pokémonok számának megoszlása: 674 és 46.

Normális és legendás pokémonok közti különbségek

Tulajdonságok

```
layout(matrix(c(1,1,1,2,4,6,3,5,7),3,3))
labels <- c('Normal','Legendary')
labels.short <- c('L','N')
boxplot(normal$Total,legendary$Total,main="Total",names = labels)
boxplot(legendary$HP,normal$HP,main="HP",names = labels.short,horizontal = TRUE)
boxplot(legendary$Attack,normal$Attack,main="Attack",names = labels.short,horizontal = TRUE)
boxplot(legendary$Defense,normal$Defense,main="Defense",names = labels.short,horizontal = TRUE)
boxplot(legendary$Speed,normal$Speed,main="Speed",names = labels.short,horizontal = TRUE)
boxplot(legendary$Sp..Atk,normal$Sp..Atk,main="Special Attack",names = labels.short,horizontal = TRUE)
boxplot(legendary$Sp..Def,normal$Sp..Def,main="Special Defense",names = labels.short,horizontal = TRUE)
```



Ez alapján azt láthatjuk, hogy többségében a legendás pokémonok jobb tulajdonsagokkal rendelkeznek a normálisakkal szemben.

Eloszlások

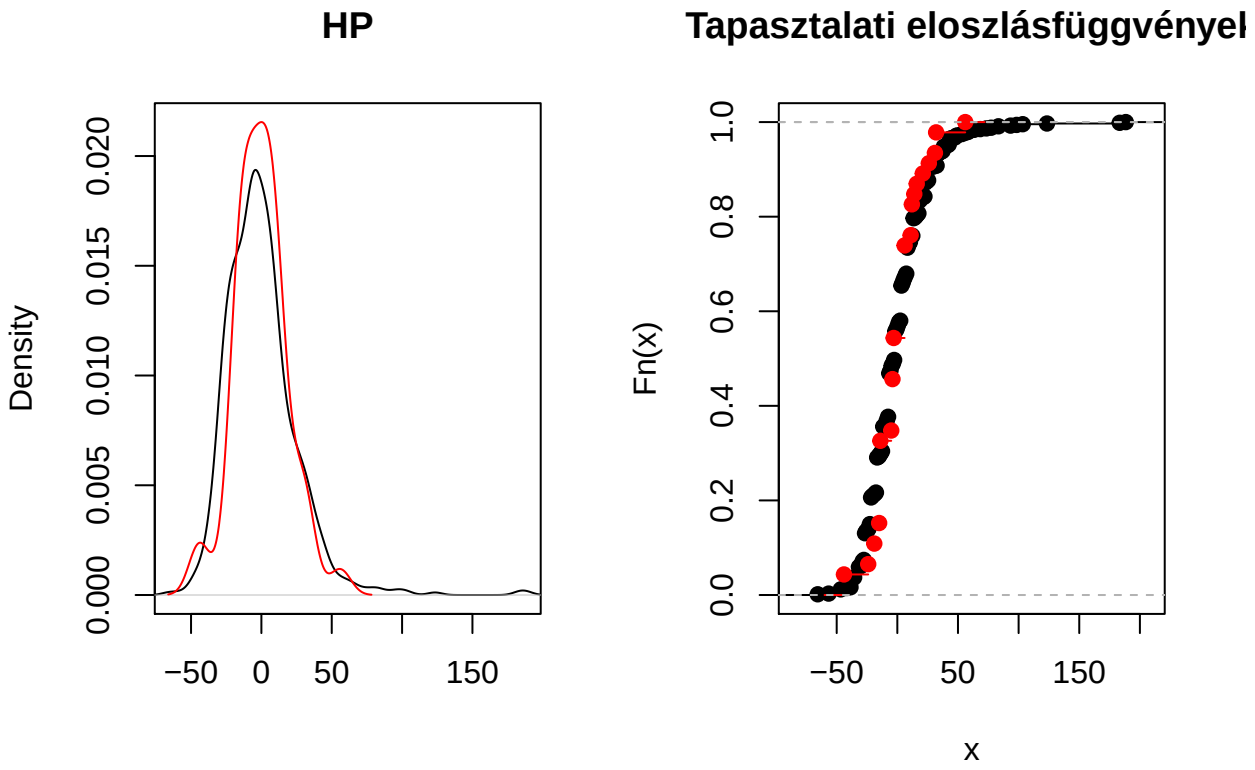
Tudjuk, hogy a legendás pokémonok tulajdonságai alapvetően magasabb értékűek a normális pokémonokénál. De vajon van-e hasonlóság abban, hogy az adott intervallumaikukon hogyan oszlanak el?

Ehhez a különböző tulajdonságok átlagát transzformáljuk valamely közös pontba, mondjuk legyen a 0. Majd kétmintás Kolmogorov-Szmirnov próbával teszteljük az eloszlások hasonlóságát, 95%-os szignifikancia szint mellett.

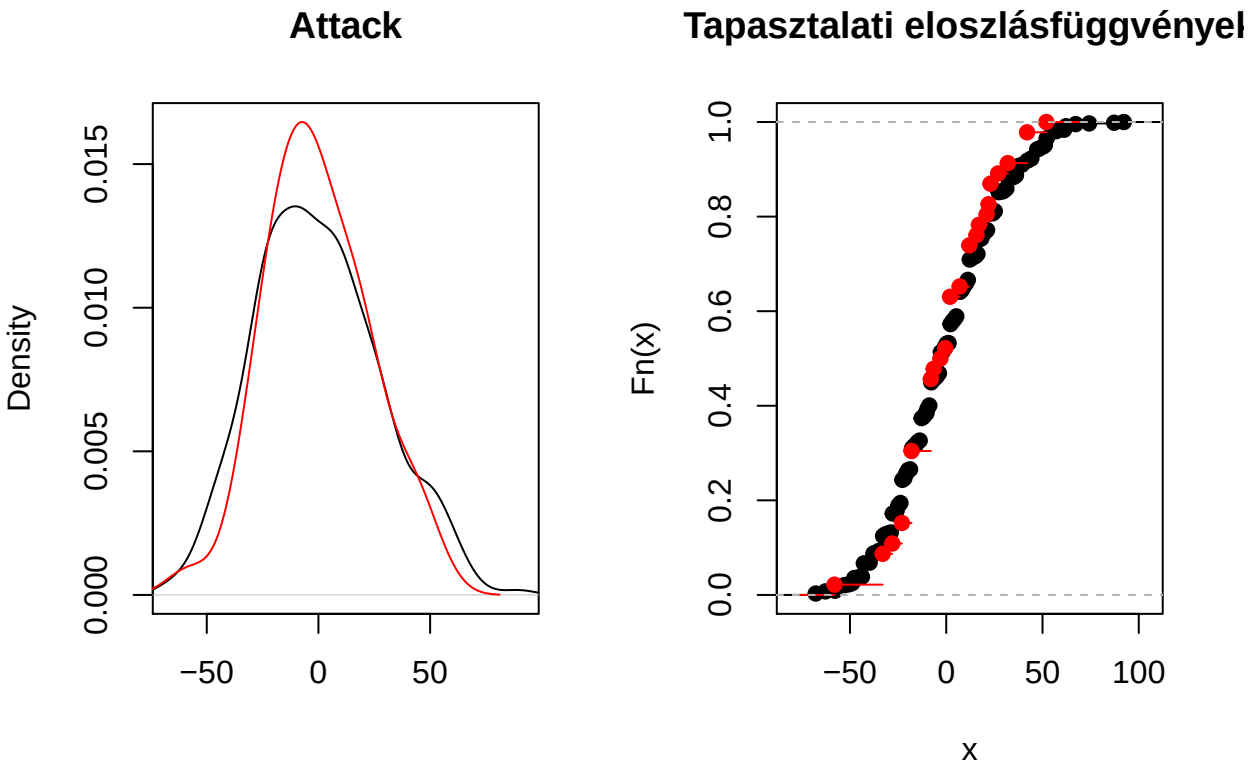
```
tests.ks <- list()

stats.names <- c('HP','Attack','Defense','Speed','Sp..Atk','Sp..Def')

for (i in stats.names) {
  layout(matrix(c(1:2),1,2,byrow = TRUE))
  tmp.norm <- normal[[i]] - mean(normal[[i]])
  tmp.legend <- legendary[[i]] - mean(legendary[[i]])
  dens.norm <- density(tmp.norm)
  dens.leg <- density(tmp.legend)
  limits.x <- c(min(min(tmp.norm),min(tmp.legend)),max(max(tmp.norm),max(tmp.legend)))
  limits.y <- c(0,max(max(dens.norm$y),max(dens.leg$y)))
  #layout(matrix(c(1,2),1,2,byrow = FALSE))
  #hist(tmp.norm,freq = FALSE,xlim = limits.x, ylim = limits.y,main = i,xlab = '')
  plot(dens.norm,xlim = limits.x, ylim = limits.y,main = i,xlab = '')
  #hist(tmp.legend,freq = FALSE,xlim = limits.x, ylim = limits.y,add=T,border = 2)
  lines(dens.leg,col = 2)
  plot(ecdf(tmp.norm),main='Tapasztalati eloszlásfüggvények')
```

```
  plot(ecdf(tmp.legend),add=TRUE,col=2)

  tests.ks[[i]] <- ks.test(tmp.norm,tmp.legend)
}
```

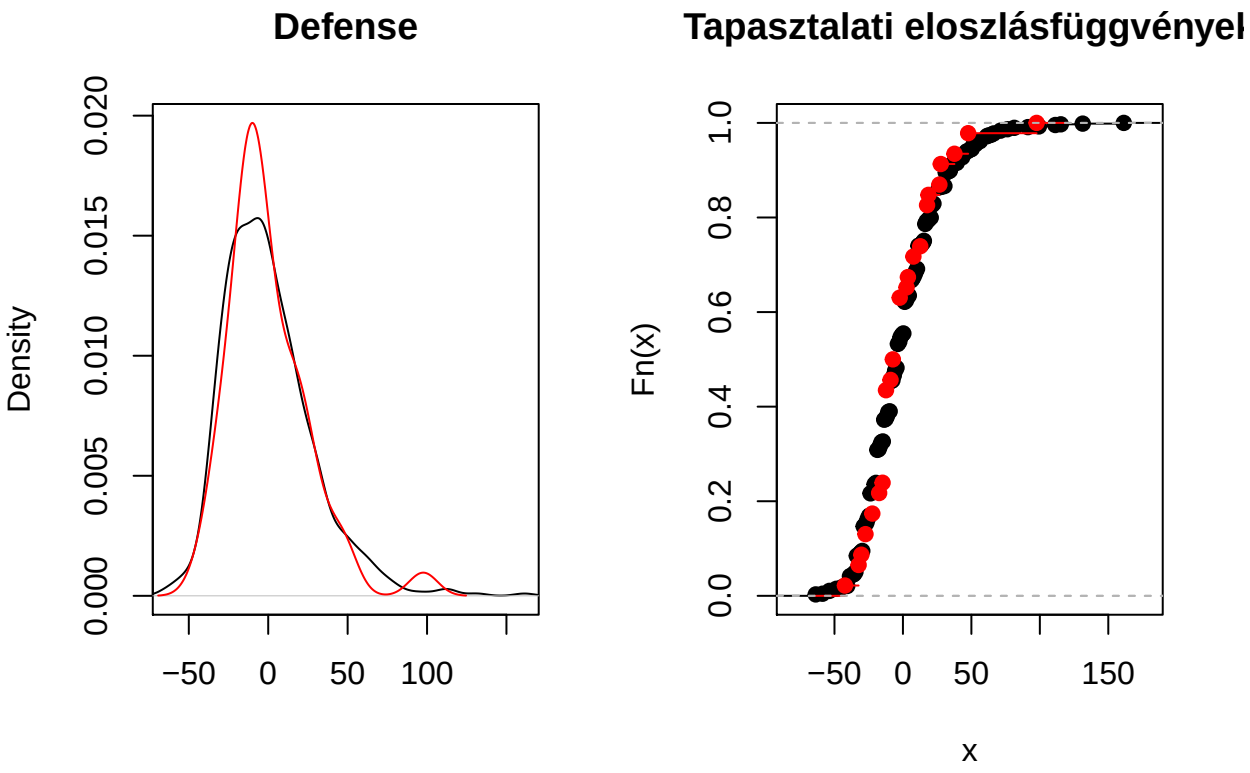
```
## Warning in ks.test(tmp.norm, tmp.legend): p-value will be approximate in
## the presence of ties
```



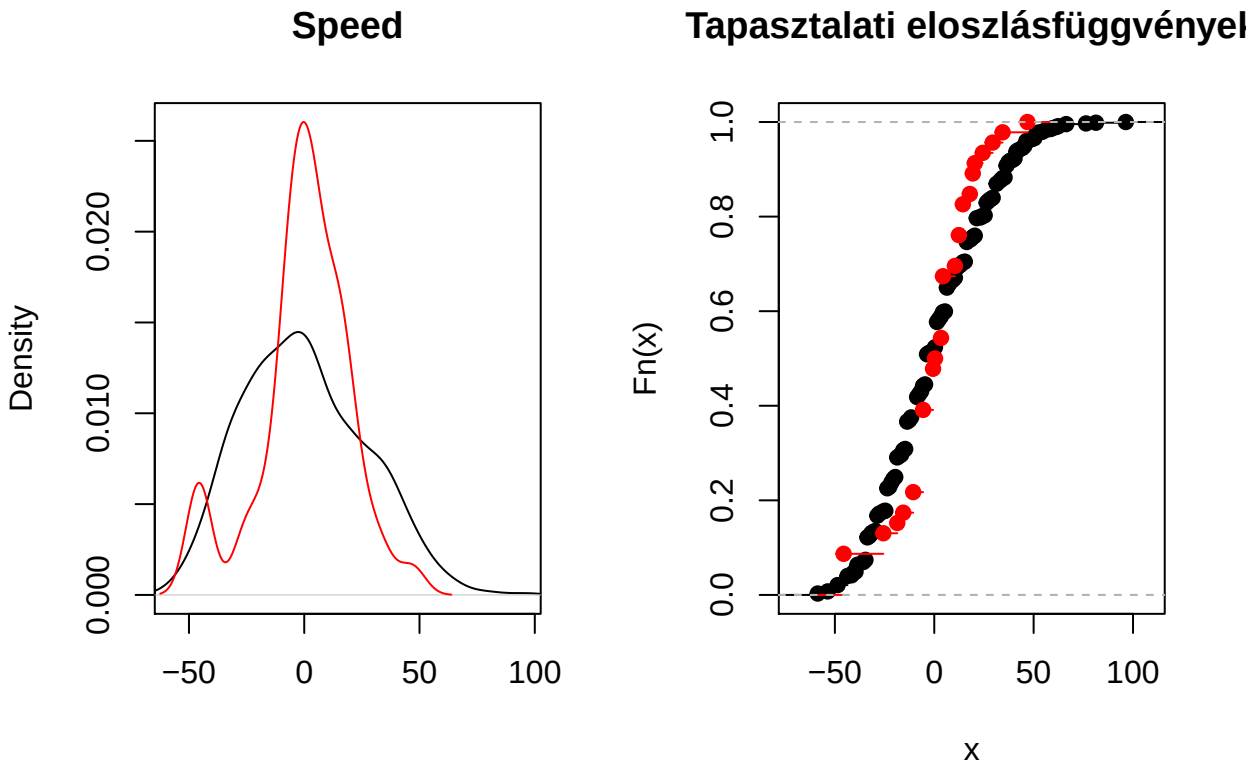

```
## Warning in ks.test(tmp.norm, tmp.legend): p-value will be approximate in
## the presence of ties
```



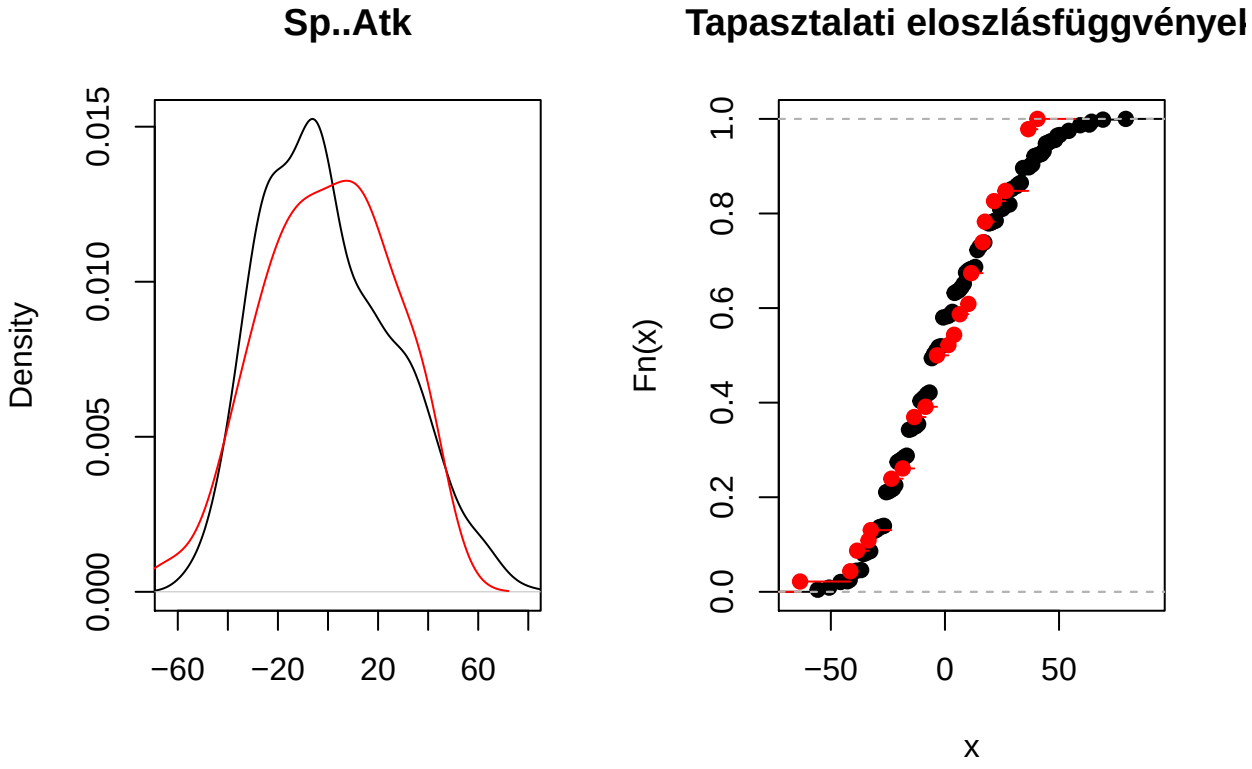

```
## Warning in ks.test(tmp.norm, tmp.legend): p-value will be approximate in
## the presence of ties
```



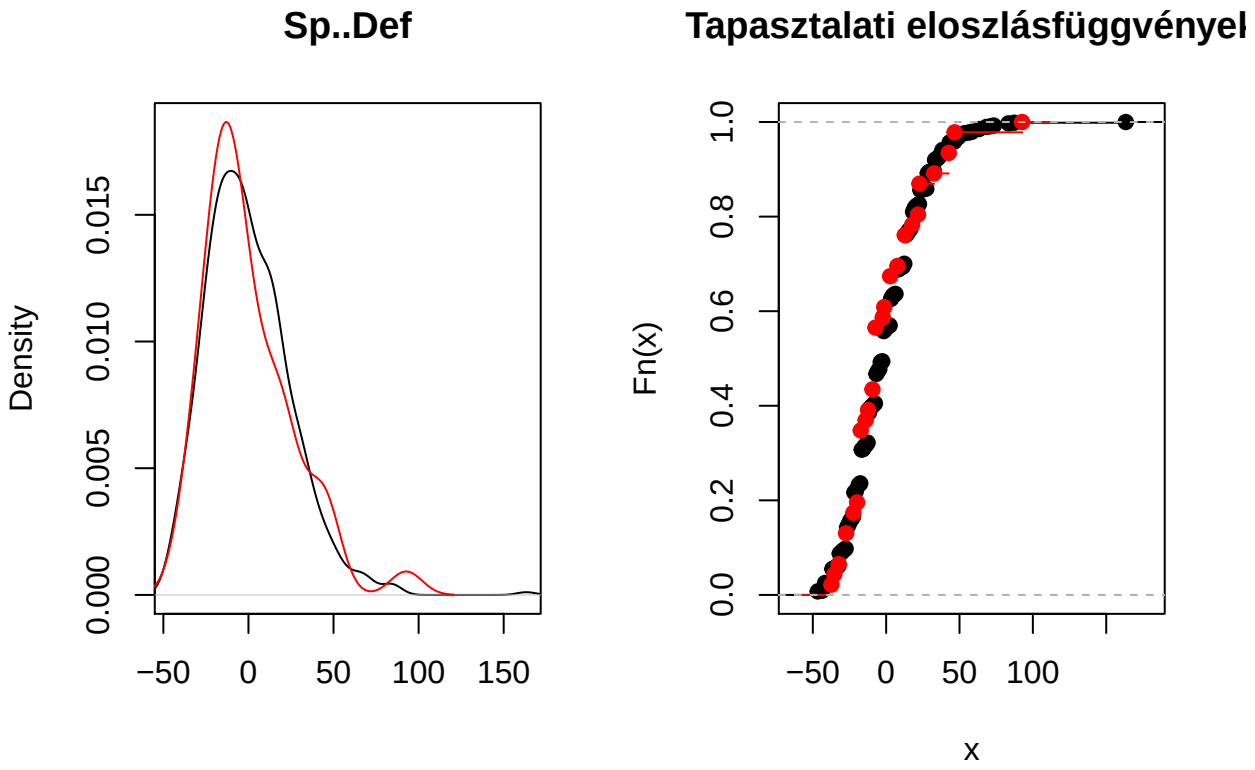

```
## Warning in ks.test(tmp.norm, tmp.legend): p-value will be approximate in
## the presence of ties
```




```
## Warning in ks.test(tmp.norm, tmp.legend): p-value will be approximate in
## the presence of ties
```




```
## Warning in ks.test(tmp.norm, tmp.legend): p-value will be approximate in
## the presence of ties
```




```
ks.value <- function(alpha,n,m) {
  sqrt(-1/2*log(alpha/2)) * sqrt((n+m)/(n*m))
}

#tests.f
alpha = 0.05
c <- ks.value(alpha,length(normal$Name),length(legendary$Name))
```

95%-os szignifikancia szint mellett a kritikus tartomány: $|D| > 0.2069615$.

- Életerő
  - A próba statisztika értéke $0.1850729 \leq 0.2069615$, tehát nem vetettük el a nullhipotézist.
- Támadás
  - A próba statisztika értéke $0.1134047 \leq 0.2069615$, tehát nem vetettük el a nullhipotézist.
- Védekezés
  - A próba statisztika értéke $0.136176 \leq 0.2069615$, tehát nem vetettük el a nullhipotézist.
- Sebesség
  - A próba statisztika értéke $0.2247452 > 0.2069615$, tehát elvetettük a nullhipotézist.
- Speciális Támadás
  - A próba statisztika értéke $0.1190814 \leq 0.2069615$, tehát nem vetettük el a nullhipotézist.
- Speciális Védekezés
  - A próba statisztika értéke $0.1601729 \leq 0.2069615$, tehát nem vetettük el a nullhipotézist.

Típusok

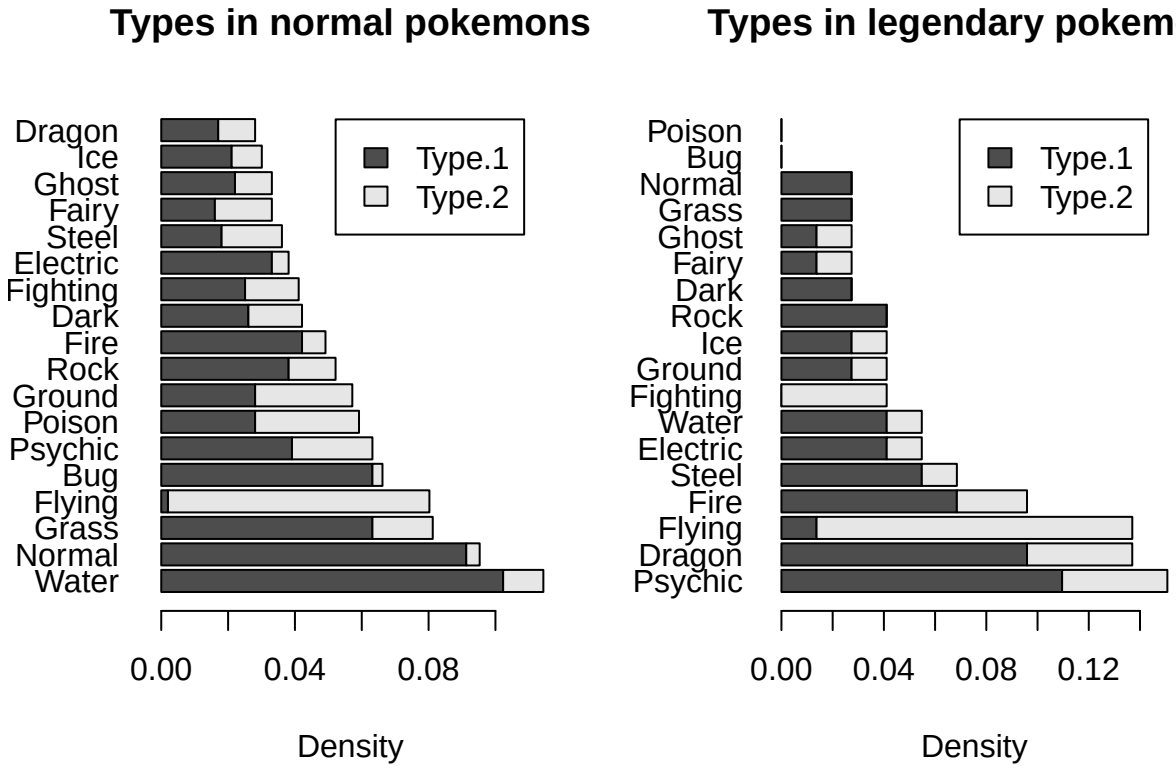
A pokémonoknak legalább egy és maximum két típusa lehet. A típusnak a támadásoknál van szerepe, a különféle képességek más más típusra máshogy hathatnak. (“It’s super effective!”) Mivel a két típussal rendelkező pokémonok mindkét típusuknak megfelelően erősebbek és gyengébbek a más típusokkal szemben, a típus szerinti összegzésbe mindkét típus szerint beleszámoltam őket. Az adatok összessége így több mint az

eredeti, de talán átfogóbb képet kaphatunk.

Másodlagos típus

Először az elsődleges és másodlagos típusok megoszlását külön a normális és a legendás pokémonok között vizsgáltam.

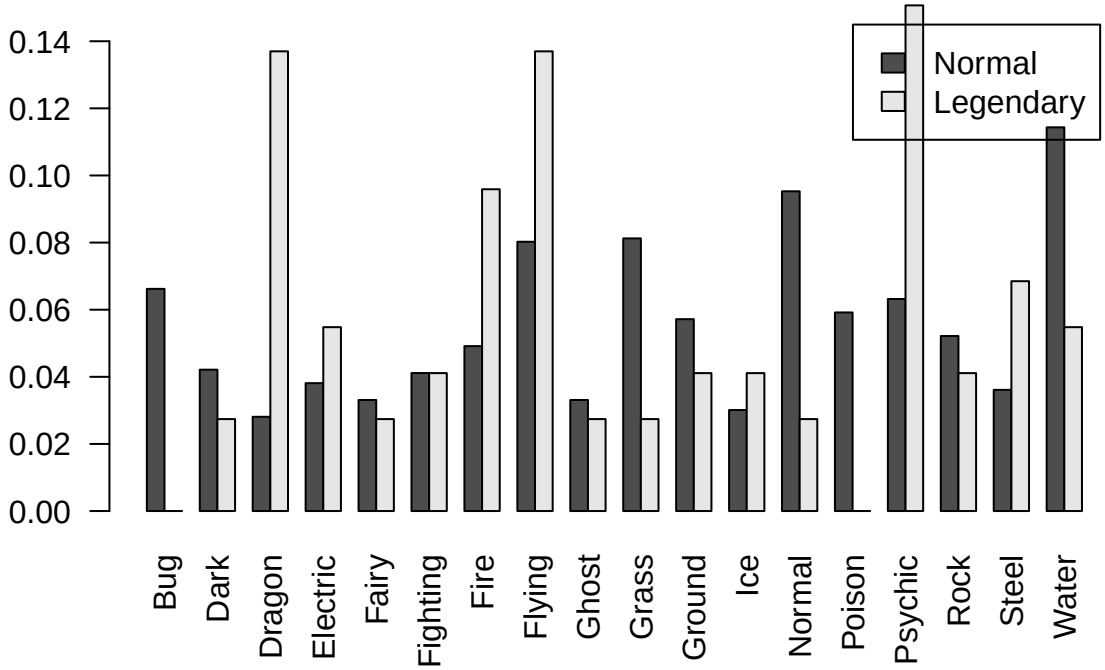
```
norm.type1.summary <- summary(normal$Type.1)
norm.type2.summary <- summary(normal$Type.2)[-1]
norm.types.summary <- t(cbind(as.matrix(norm.type1.summary),as.matrix(norm.type2.summary)))
norm.types.summary.normaled <- norm.types.summary/sum(norm.types.summary)
norm.types.summary.normaled.ordered <- norm.types.summary.normaled[,order(colSums(norm.types.summary.no
leg.type1.summary <- summary(legendary$Type.1)
leg.type2.summary <- summary(legendary$Type.2)[-1]
leg.types.summary <- t(cbind(as.matrix(leg.type1.summary),as.matrix(leg.type2.summary)))
leg.types.summary.normaled <- leg.types.summary/sum(leg.types.summary)
leg.types.summary.normaled.ordered <- leg.types.summary.normaled[,order(colSums(leg.types.summary.normal
types.summary <- cbind(colSums(norm.types.summary),colSums(leg.types.summary))
types.summary.normaled <- cbind(types.summary[,1]/sum(types.summary[,1]),types.summary[,2]/sum(types.su

maxval <- max(max(colSums(norm.types.summary.normaled)),max(colSums(leg.types.summary.normaled)))
maxval <- c(0,maxval)

label.x <- 'Density'
legendtext.1 <- c('Type.1','Type.2')

layout(matrix(c(1,2),1,2,byrow = TRUE))
barplot(norm.types.summary.normaled.ordered,las=1,horiz = TRUE,legend.text = legendtext.1,main = 'Types
barplot(leg.types.summary.normaled.ordered,las=1,horiz = TRUE,xlim = maxval,legend.text = legendtext.1,
```



```
num.water <- sum(norm.types.summary[,colnames(norm.types.summary)=='Water'])
num.normal <- sum(norm.types.summary[,colnames(norm.types.summary)=='Normal'])
num.sum <- sum(norm.types.summary)
```

- A repülő típus főleg másodlagosként szerepel
- Normális pokémonoknál a legyakoribb típus a víz, de ez 114 darabot jelent, míg a második leggyakoribb, a "Normal" 95 darab (az 997-ből), ez pedig nem tekinthető szignifikáns különbségnek (1.9057172%).
- Legendás pokémonoknál jellemzőbbnek tűnik a második típus megléte.
  - A legendás pokémonoknak 58.6956522%-a,
  - Ezzel szemben a normálisoknak 47.9228487%-a rendelkezik másodlagos típussal

```
layout(matrix(1,1,1))
barplot(t(types.summary.normaled),beside = TRUE,las = 2,legend.text = labels)
```



Észrevehető, hogy legendás pokémonok gyakrabban sárkány, repülő illetve psychic típusúak.

Generáció szerinti bontás

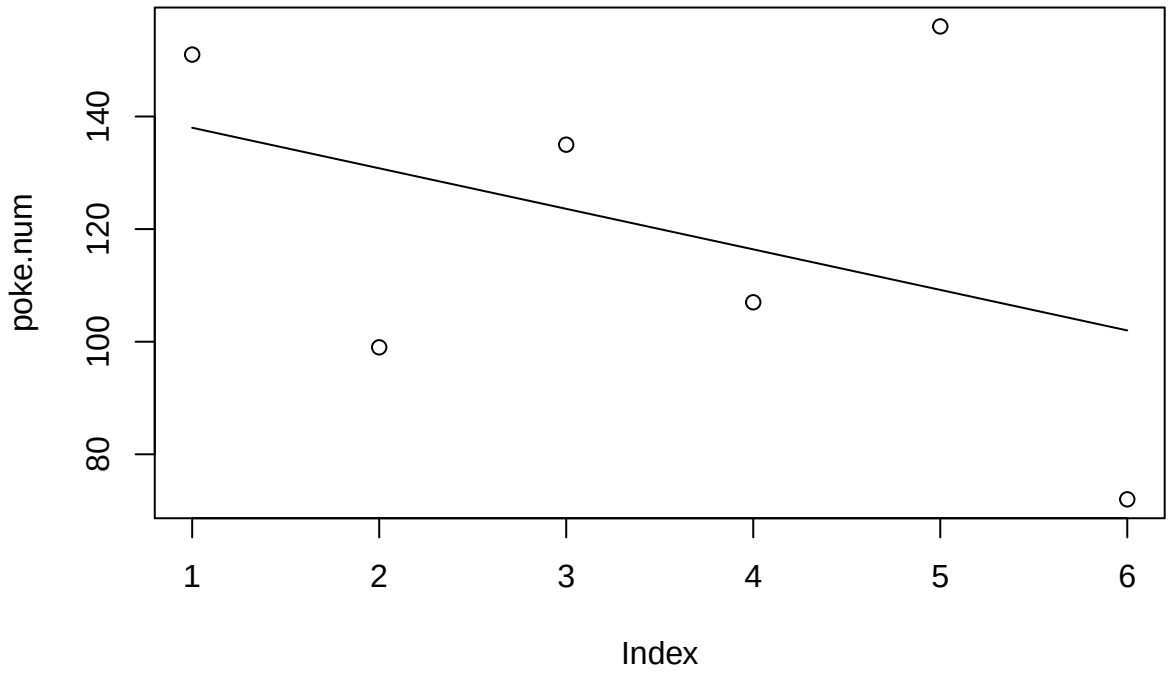
```
data.gens <- split(data,data$Generation)
normal.gens <- split(normal,normal$Generation)
legendary.gens <- split(legendary,legendary$Generation)
```

Pokémonok száma

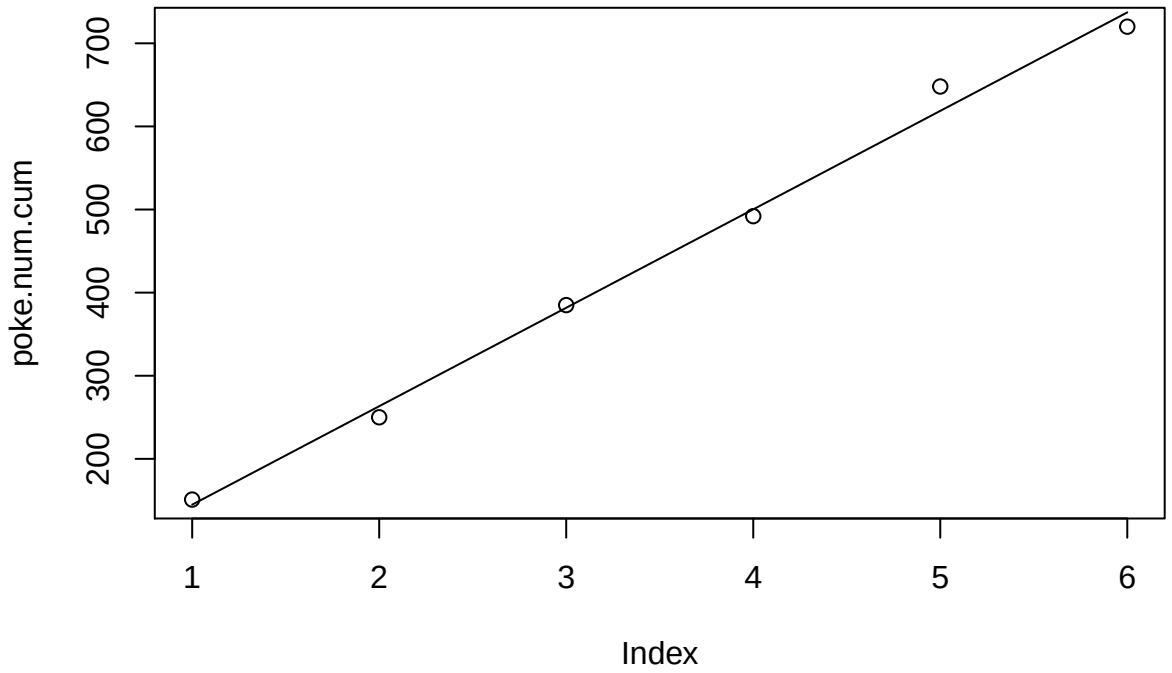
Megvizsgáltam, hogy a generációkban hogyan oszlik meg az új pokémonok száma, és erre lineáris modell illesztésével előrejelezni, hogy a következő

```
poke.num <- rep(0,6)
poke.num.cum <- rep(0,6)
gen.list <- c(1:length(data.gens))
```

```r
for (i in gen.list) {
  poke.num[i] <- length(data.gens[[i]]$Name)
  poke.num.cum[i] <- sum(poke.num[1:i])
}

regr <- function(coefs,x) coefs[1] + x * coefs[2]

plot(poke.num)
num.model <- lm(poke.num~gen.list)
lines(num.model$fitted.values)
```




```r
#lines(regr(num.model$coefficients,c(1:6)))

plot(poke.num.cum)
num.cum.model <- lm(poke.num.cum~gen.list)
lines(num.cum.model$fitted.values)
```



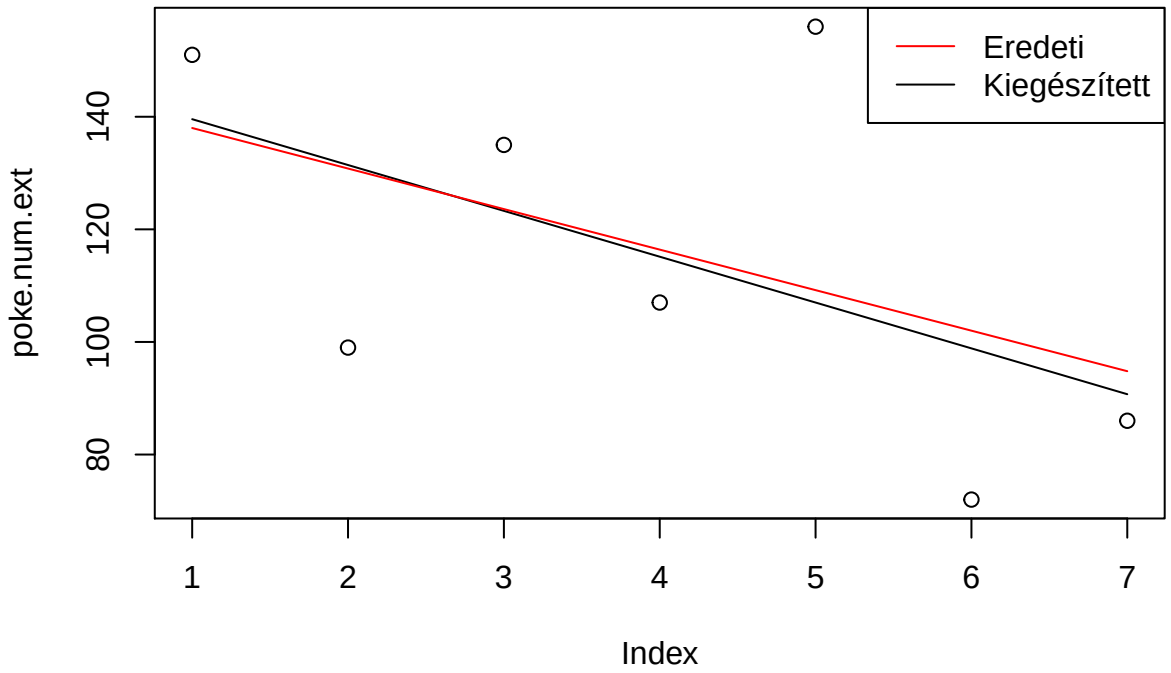

```
#regr <- function(coefs,x) sum(c(1,x)*coefs)
gen7.num <- regr(num.model$coefficients,7)
gen7.num.cum <- regr(num.cum.model$coefficients,7)
```

A modell alapján a 7. generációban 94.8 új pokémon lesz, azaz összesen 814.8. A kumulált modell szerint a 7. generációban összesen 855.6 darab ($\Delta = 40.8$). A valódi 7. generációban ezzel szemben 86 új pokémon volt, így lett összesen 807.

Tehát ha nem a kumulált pokémon számot néztük, pontosabb becslést kapunk.

Ha a modellünket kiegészítjük a 7. generáció pokémonjainak számával a 8. generációra az alábbi becslést kapjuk.

```
gen7.num.exact <- 86
gen.list.ext <- c(1:7)
poke.num.ext <- c(poke.num,gen7.num.exact)
plot(poke.num.ext)
legend('topright', legend = c('Eredeti','Kiegészített'), col=c(2,1),lty = 1)
num.ext.model <- lm(poke.num.ext~gen.list.ext)
lines(num.ext.model$fitted.values)
lines(regr(num.model$coefficients,c(1:7)),col=2)
```



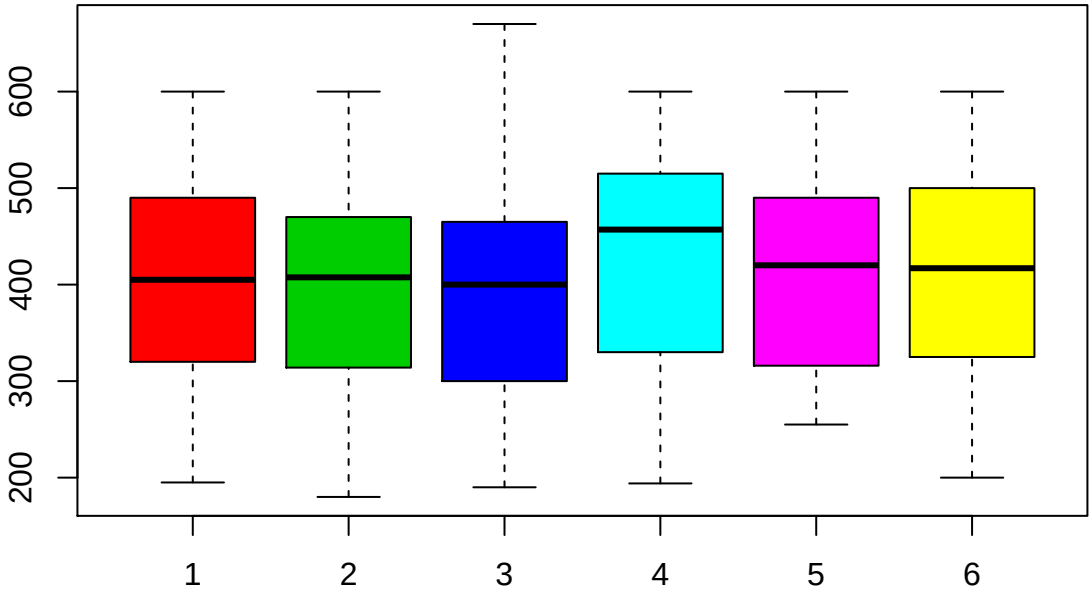

```
gen8.num <- c(0,0)
gen8.num[1] <- regr(num.model$coefficients,8)
gen8.num[2] <- regr(num.ext.model$coefficients,8)
#num.model$coefficients - num.ext.model$coefficients
```

Ez alapján a modell alapján a becslés a 8. generáció új pokémonjainak számára 82.5714286, míg az előző modell alapján 87.6 ($\Delta = 5.0285714$).

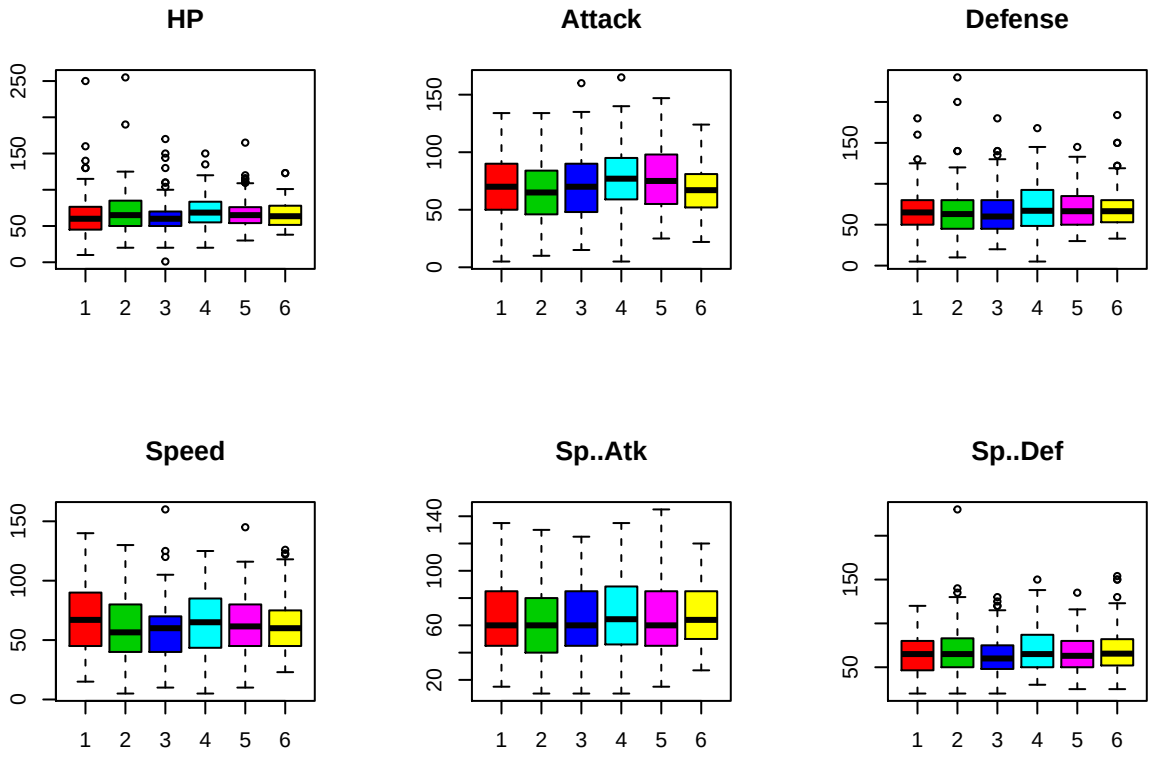
Tulajdonságok

Generációkra lebontva a legendás pokémonok száma elenyésző, így őket nem vettem bele a vizsgálatba a továbbiakban.

```
#gen1 <- list(both = data.gens$'1', normal = normal.gens$'1', legendary = legendary.gens$'1')
gen1 <- normal.gens$'1'
gen2 <- normal.gens$'2'
gen3 <- normal.gens$'3'
gen4 <- normal.gens$'4'
gen5 <- normal.gens$'5'
gen6 <- normal.gens$'6'
boxplot(gen1$Total,gen2$Total,gen3$Total,gen4$Total,gen5$Total,gen6$Total,main='Generációk össz tulajdo
```




```
layout(matrix(c(1:6),2,3,byrow = TRUE))
for (i in stats.names) {
  boxplot(gen1[[i]],gen2[[i]],gen3[[i]],gen4[[i]],gen5[[i]],gen6[[i]],main=i,names = names(data.gens),c
}
```

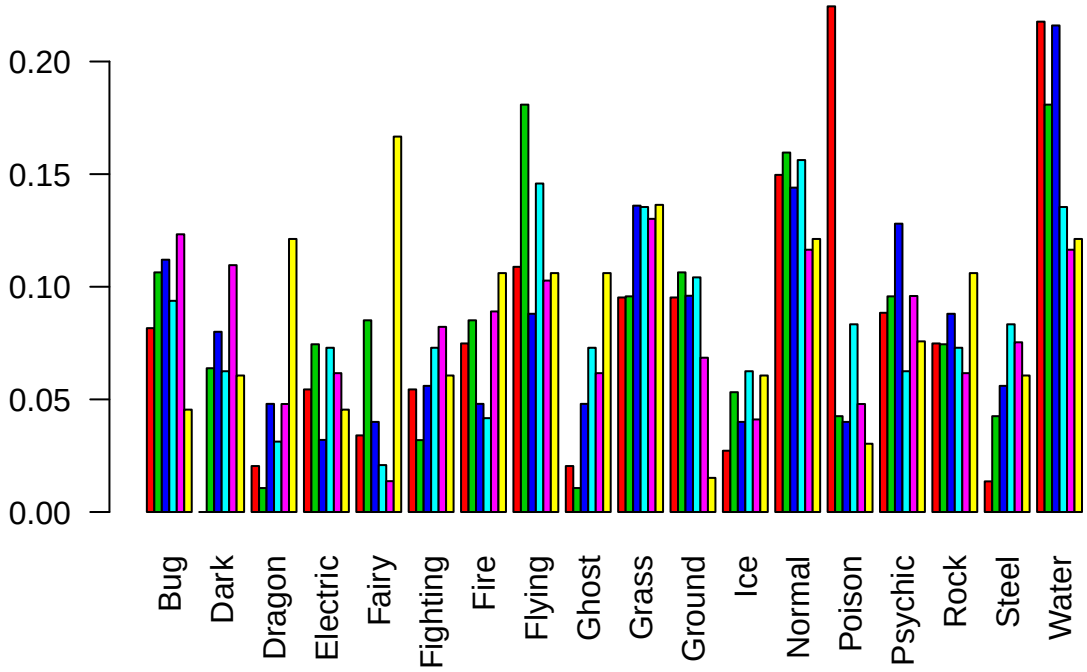


Innen azt vehetjük észre, hogy a negyedik generációban összességében erősebb pokémonok érkeztek.

Típusok

Itt arra voltam kíváncsi, hogy vannak-e különböző típus trendek a különböző generációkban.

```
gens.types.summary <- NULL
gens.types.summary.normaled <- NULL

for (gen in normal.gens) {
  #print(i$Type.1)

gen.type1.summary <- summary(gen$Type.1)
gen.type2.summary <- summary(gen$Type.2)[-1]
gen.types.summary <- t(cbind(as.matrix(gen.type1.summary),as.matrix(gen.type2.summary)))
types.sum <- colSums(gen.types.summary)
types.summary.normaled <- colSums(gen.types.summary) / nrow(gen)
gens.types.summary <- cbind(gens.types.summary,types.sum)
gens.types.summary.normaled <- cbind(gens.types.summary.normaled,types.summary.normaled)

#barplot(types.summary)
}
#gens.types.summary <- t(gens.types.summary)
gens.types.summary.normaled <- t(gens.types.summary.normaled)
#barplot(gens.types.summary,horiz = FALSE,las=2,beside = TRUE)
barplot(gens.types.summary.normaled,horiz = FALSE,las=2,beside = TRUE,col=c(2:7))
```



- Az első generációban a méreg és a víz típusú pokémonok a gyakoriak, de a méreg típus csak az első generációban volt ilyen meghatározó.
- A második generációban a repülő pokémonok száma kiugró.
- A hatodik generációban pedig a “Fairy” típus örvendett nagy népszerűségnek.

Próba

Legyen a nullhipotézisünk, hogy a pokémon típusa nem függ attól hogy hanyadik generációs.

```r
alpha <- 0.05
gens.types.summary
```

```
##          types.sum types.sum types.sum types.sum types.sum types.sum
## Bug             12        10        14         9        18         3
## Dark             0         6        10         6        16         4
## Dragon           3         1         6         3         7         8
## Electric         8         7         4         7         9         3
## Fairy            5         8         5         2         2        11
## Fighting         8         3         7         7        12         4
## Fire            11         8         6         4        13         7
## Flying          16        17        11        14        15         7
## Ghost            3         1         6         7         9         7
## Grass           14         9        17        13        19         9
## Ground          14        10        12        10        10         1
## Ice              4         5         5         6         6         4
## Normal          22        15        18        15        17         8
## Poison          33         4         5         8         7         2
## Psychic         13         9        16         6        14         5
## Rock            11         7        11         7         9         7
## Steel            2         4         7         8        11         4
## Water           32        17        27        13        17         8
```

```r
res <- chisq.test(gens.types.summary)
```

```
## Warning in chisq.test(gens.types.summary): Chi-squared approximation may be
## incorrect
```

```r
crit <- qchisq(alpha,res$parameter)
res
```

```
##
##  Pearson's Chi-squared test
##
## data:  gens.types.summary
## X-squared = 157.73, df = 85, p-value = 2.809e-06
```

A próbához tartozó kritikus tartományunk 5%-os szignifikancia szint mellett : $\chi^2 > 64.7493958$. A próba eredménye $157.7262994 > 64.7493958$, tehát a nullhipotézist elvetjük, és a pokémon típusa függ attól, hogy hanyadik generációs.

A dokumentumról

Ez a dokumentum R Studio-ban készült, a forrása egy `.Rmd` (R Markdown) fájl (mellékelve). A dokumentumban szereplő kódrészletek a generálásakor kiértékelődtek, és azok kimenete látszik. A csatolt `.r` fájlban a dokumentumban szereplő kódrészletek vannak összegyűjtve.