Analiza UFC borbi

2023-01-15

Prvi pogled na podatke

Učitajmo skup podataka i pogledajmo koje značajke su dostupne.

```
data=read.csv("UFC.csv")

names(data)
```

```
##   [1] "Winner"                          "title_bout"
##   [3] "B_avg_KD"                        "B_avg_opp_KD"
##   [5] "B_avg_SIG_STR_pct"               "B_avg_opp_SIG_STR_pct"
##   [7] "B_avg_TD_pct"                    "B_avg_opp_TD_pct"
##   [9] "B_avg_SUB_ATT"                   "B_avg_opp_SUB_ATT"
##  [11] "B_avg_REV"                       "B_avg_opp_REV"
##  [13] "B_avg_SIG_STR_att"               "B_avg_SIG_STR_landed"
##  [15] "B_avg_opp_SIG_STR_att"           "B_avg_opp_SIG_STR_landed"
##  [17] "B_avg_TOTAL_STR_att"             "B_avg_TOTAL_STR_landed"
##  [19] "B_avg_opp_TOTAL_STR_att"         "B_avg_opp_TOTAL_STR_landed"
##  [21] "B_avg_TD_att"                    "B_avg_TD_landed"
##  [23] "B_avg_opp_TD_att"                "B_avg_opp_TD_landed"
##  [25] "B_avg_HEAD_att"                  "B_avg_HEAD_landed"
##  [27] "B_avg_opp_HEAD_att"              "B_avg_opp_HEAD_landed"
##  [29] "B_avg_BODY_att"                  "B_avg_BODY_landed"
##  [31] "B_avg_opp_BODY_att"              "B_avg_opp_BODY_landed"
##  [33] "B_avg_LEG_att"                   "B_avg_LEG_landed"
##  [35] "B_avg_opp_LEG_att"               "B_avg_opp_LEG_landed"
##  [37] "B_avg_DISTANCE_att"              "B_avg_DISTANCE_landed"
##  [39] "B_avg_opp_DISTANCE_att"          "B_avg_opp_DISTANCE_landed"
##  [41] "B_avg_CLINCH_att"                "B_avg_CLINCH_landed"
##  [43] "B_avg_opp_CLINCH_att"            "B_avg_opp_CLINCH_landed"
##  [45] "B_avg_GROUND_att"                "B_avg_GROUND_landed"
##  [47] "B_avg_opp_GROUND_att"            "B_avg_opp_GROUND_landed"
##  [49] "B_avg_CTRL_time.seconds."        "B_avg_opp_CTRL_time.seconds."
##  [51] "B_total_time_fought.seconds."    "B_total_rounds_fought"
##  [53] "B_total_title_bouts"             "B_current_win_streak"
##  [55] "B_current_lose_streak"           "B_longest_win_streak"
##  [57] "B_wins"                          "B_losses"
##  [59] "B_draw"                          "B_win_by_Decision_Majority"
##  [61] "B_win_by_Decision_Split"         "B_win_by_Decision_Unanimous"
##  [63] "B_win_by_KO.TKO"                 "B_win_by_Submission"
##  [65] "B_win_by_TKO_Doctor_Stoppage"    "B_Height_cms"
##  [67] "B_Reach_cms"                     "B_Weight_lbs"
##  [69] "R_avg_KD"                        "R_avg_opp_KD"
##  [71] "R_avg_SIG_STR_pct"               "R_avg_opp_SIG_STR_pct"
```

```
##  [73] "R_avg_TD_pct"                    "R_avg_opp_TD_pct"
##  [75] "R_avg_SUB_ATT"                   "R_avg_opp_SUB_ATT"
##  [77] "R_avg_REV"                       "R_avg_opp_REV"
##  [79] "R_avg_SIG_STR_att"               "R_avg_SIG_STR_landed"
##  [81] "R_avg_opp_SIG_STR_att"           "R_avg_opp_SIG_STR_landed"
##  [83] "R_avg_TOTAL_STR_att"             "R_avg_TOTAL_STR_landed"
##  [85] "R_avg_opp_TOTAL_STR_att"         "R_avg_opp_TOTAL_STR_landed"
##  [87] "R_avg_TD_att"                    "R_avg_TD_landed"
##  [89] "R_avg_opp_TD_att"                "R_avg_opp_TD_landed"
##  [91] "R_avg_HEAD_att"                  "R_avg_HEAD_landed"
##  [93] "R_avg_opp_HEAD_att"              "R_avg_opp_HEAD_landed"
##  [95] "R_avg_BODY_att"                  "R_avg_BODY_landed"
##  [97] "R_avg_opp_BODY_att"              "R_avg_opp_BODY_landed"
##  [99] "R_avg_LEG_att"                   "R_avg_LEG_landed"
## [101] "R_avg_opp_LEG_att"               "R_avg_opp_LEG_landed"
## [103] "R_avg_DISTANCE_att"              "R_avg_DISTANCE_landed"
## [105] "R_avg_opp_DISTANCE_att"          "R_avg_opp_DISTANCE_landed"
## [107] "R_avg_CLINCH_att"                "R_avg_CLINCH_landed"
## [109] "R_avg_opp_CLINCH_att"            "R_avg_opp_CLINCH_landed"
## [111] "R_avg_GROUND_att"                "R_avg_GROUND_landed"
## [113] "R_avg_opp_GROUND_att"            "R_avg_opp_GROUND_landed"
## [115] "R_avg_CTRL_time.seconds."        "R_avg_opp_CTRL_time.seconds."
## [117] "R_total_time_fought.seconds."    "R_total_rounds_fought"
## [119] "R_total_title_bouts"             "R_current_win_streak"
## [121] "R_current_lose_streak"           "R_longest_win_streak"
## [123] "R_wins"                          "R_losses"
## [125] "R_draw"                          "R_win_by_Decision_Majority"
## [127] "R_win_by_Decision_Split"         "R_win_by_Decision_Unanimous"
## [129] "R_win_by_KO.TKO"                 "R_win_by_Submission"
## [131] "R_win_by_TKO_Doctor_Stoppage"    "R_Height_cms"
## [133] "R_Reach_cms"                     "R_Weight_lbs"
## [135] "B_age"                           "R_age"
## [137] "weight_class_Bantamweight"       "weight_class_CatchWeight"
## [139] "weight_class_Featherweight"      "weight_class_Flyweight"
## [141] "weight_class_Heavyweight"        "weight_class_LightHeavyweight"
## [143] "weight_class_Lightweight"        "weight_class_Middleweight"
## [145] "weight_class_OpenWeight"         "weight_class_Welterweight"
## [147] "weight_class_WomenBantamweight"  "weight_class_WomenFeatherweight"
## [149] "weight_class_WomenFlyweight"     "weight_class_WomenStrawweight"
## [151] "B_Stance_Open.Stance"            "B_Stance_Orthodox"
## [153] "B_Stance_Sideways"               "B_Stance_Southpaw"
## [155] "B_Stance_Switch"                 "R_Stance_Open.Stance"
## [157] "R_Stance_Orthodox"               "R_Stance_Sideways"
## [159] "R_Stance_Southpaw"               "R_Stance_Switch"
```

Pogledajmo koliko zapisa i koliko značajki ima u skupu podataka.

```
dim(data)
```

```
## [1] 5902  160
```

Pitanje 1.

Možemo li očekivati završetak borbe nokautom ovisno o razlici u dužini ruku izmedu boraca?

Učitavamo podatke i ostavljamo stupce koji će nam biti potrebni za testiranje ("B_Reach_cms","R_Reach_cms", "win_by"). Također stvaramo stupac razlika u kojem će biti pospremljena razlika u dužini ruku između boraca u toj borbi.

```
readfile = read.csv("combined.csv")
df = readfile[c("Winner","B_Reach_cms","R_Reach_cms", "win_by")]
razlika = df$B_Reach_cms - df$R_Reach_cms
df$razlika = razlika
```

Radimo deskriptivnu statistiku kako bi se bolje upoznali s podatcima.

```
names(df)
```

```
## [1] "Winner"      "B_Reach_cms" "R_Reach_cms" "win_by"      "razlika"
```

```
dim(df)
```

```
## [1] 5902    5
```

```
summary(df)
```

```
##     Winner           B_Reach_cms     R_Reach_cms       win_by
##  Length:5902        Min.   :147.3   Min.   :152.4   Length:5902
##  Class :character   1st Qu.:177.8   1st Qu.:177.8   Class :character
##  Mode  :character   Median :182.9   Median :182.9   Mode  :character
##                     Mean   :182.8   Mean   :183.5
##                     3rd Qu.:190.5   3rd Qu.:190.5
##                     Max.   :213.4   Max.   :213.4
##     razlika
##  Min.   :-33.0200
##  1st Qu.: -5.0800
##  Median :  0.0000
##  Mean   : -0.6227
##  3rd Qu.:  5.0800
##  Max.   : 27.9400
```

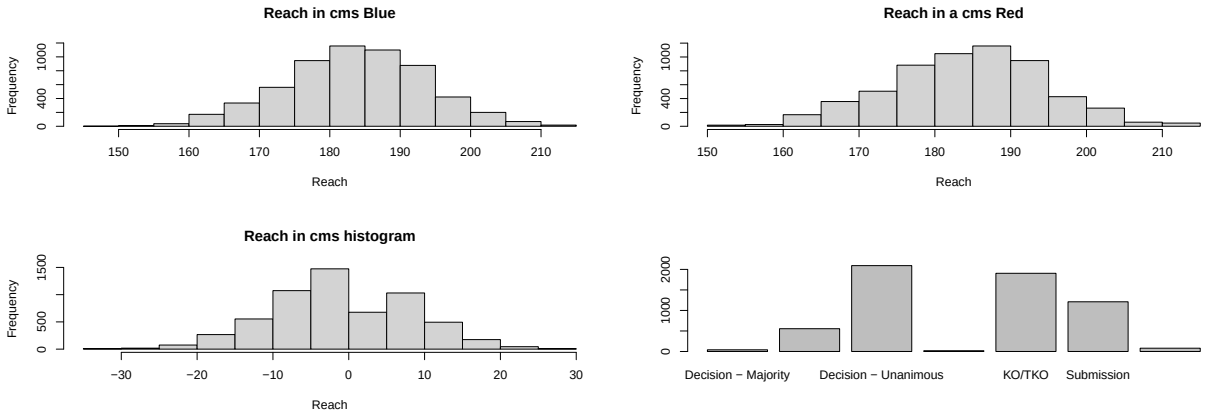
Tražimo nedostajuće vrijednosti ako ih ima.

```
for (col_name in names(df)){
  if (sum(is.na(df[,col_name])) >= 0){
    cat('Ukupno nedostajućih vrijednosti za varijablu ',col_name, ': ', sum(is.na(df[,col_name])),'\n')
  }
}
```

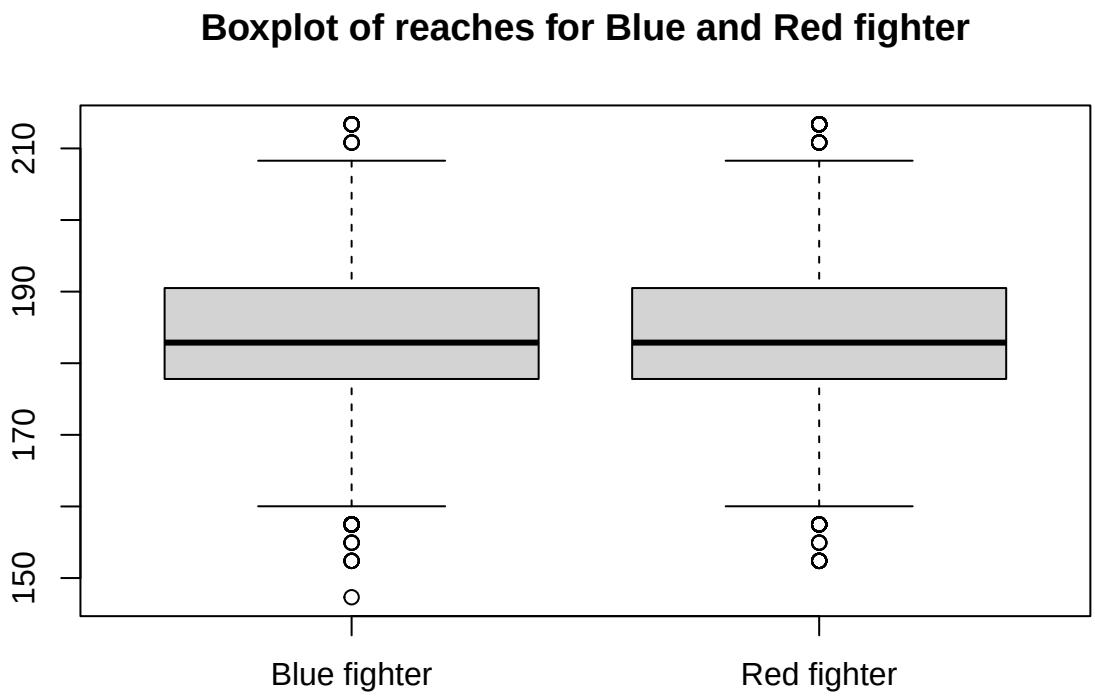
```
## Ukupno nedostajućih vrijednosti za varijablu  Winner :  0
## Ukupno nedostajućih vrijednosti za varijablu  B_Reach_cms :  0
## Ukupno nedostajućih vrijednosti za varijablu  R_Reach_cms :  0
## Ukupno nedostajućih vrijednosti za varijablu  win_by :  0
## Ukupno nedostajućih vrijednosti za varijablu  razlika :  0
```

Preko histograma vidimo normalnu razdiobu podataka za duljinu ruku pojedinog borca i međusobnu razliku u duljinama ruku.

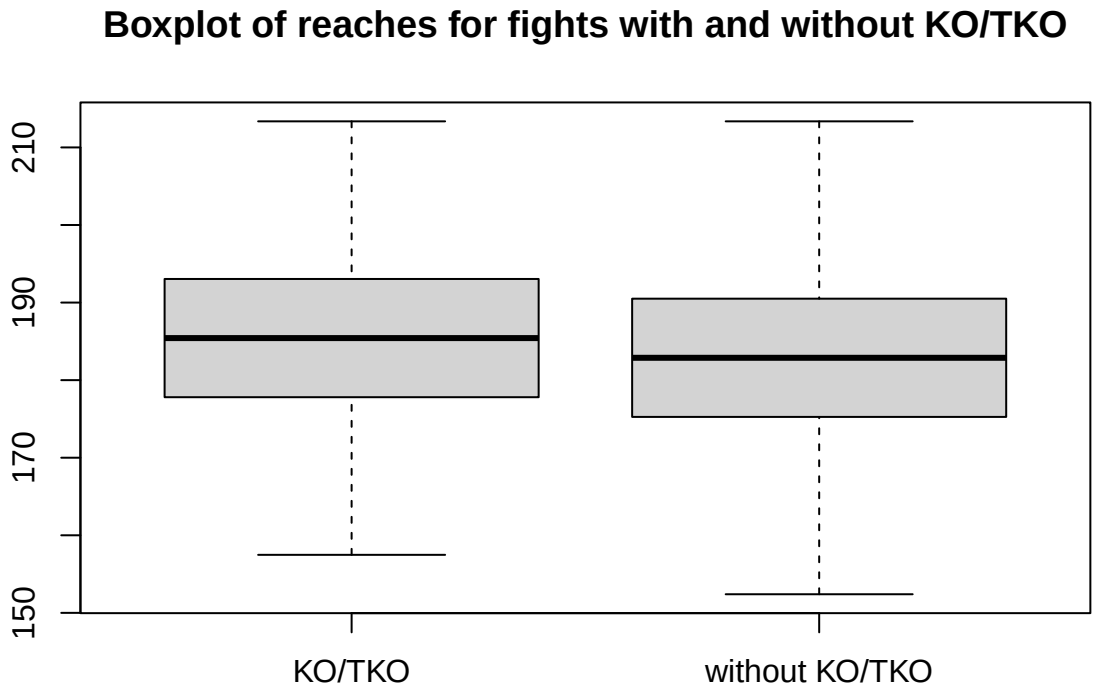
```
par(mfrow=c(2,2))

hist(df$B_Reach_cms,main='Reach in cms Blue', xlab='Reach', ylab='Frequency')
hist(df$R_Reach_cms,main='Reach in a cms Red', xlab='Reach', ylab='Frequency')
hist(df$razlika,main='Reach in cms histogram', xlab='Reach', ylab='Frequency')
barplot(table(df$win_by))
```




```
par=(mfrow=c(1,2))
boxplot(df$B_Reach_cms, df$R_Reach_cms,
        names = c('Blue fighter','Red fighter'),
        main='Boxplot of reaches for Blue and Red fighter')
```




```
boxplot(df[df$win_by == "KO/TKO",]$B_Reach_cms, df[df$win_by != "KO/TKO",]$R_Reach_cms,
        names = c('KO/TKO','without KO/TKO'),
        main='Boxplot of reaches for fights with and without KO/TKO')
```



Preko drugog boxplota primjećujemo kako je srednja vrijednost, ujedno sa Q1 i Q3 u borbama koje su završile sa KO/TKO malo veća nego u ostalim borbama. Provest ćemo test kako bi utvrdili je li razlika značajna.

Testiranje jednakosti srednjih vrijednosti dvije populacije

Kako bi proveli test, podatke ćemo podijeliti na brobe koje su završile sa KO/TKO i ostale borbe.

```
ko = df[df$win_by == "KO/TKO",]

not_ko = df[df$win_by != "KO/TKO",]
```

Hipoteze tada glase:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 < \mu_2 \quad , \quad \mu_1 > \mu_2 \quad , \quad \mu_1 \neq \mu_2$$

Gdje je

$$\mu_1$$

srednja vrijednost duljine ruku u borbama završenim nokautom i

$$\mu_2$$

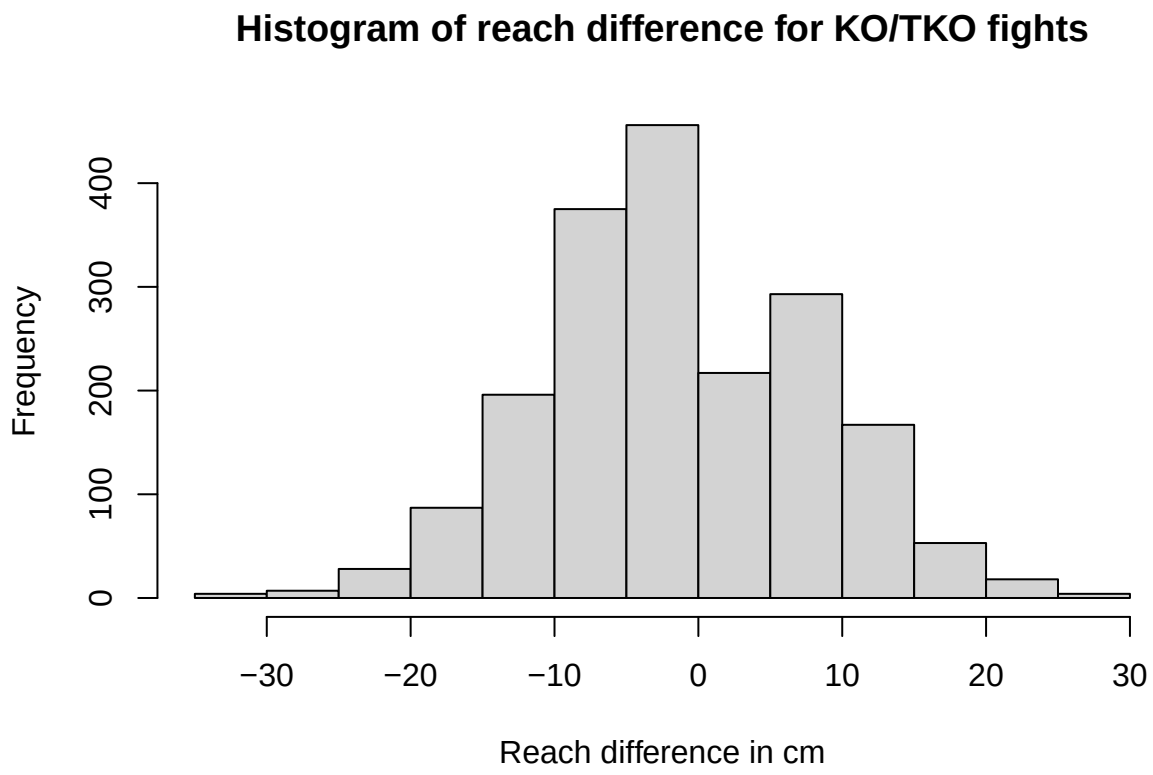
srednja vrijednost u borbama koje nisu završile nokautom.

Test o jednakosti srednjih vrijednosti dvije populacije u R-u je implementiran u funkciji `t.test()`.

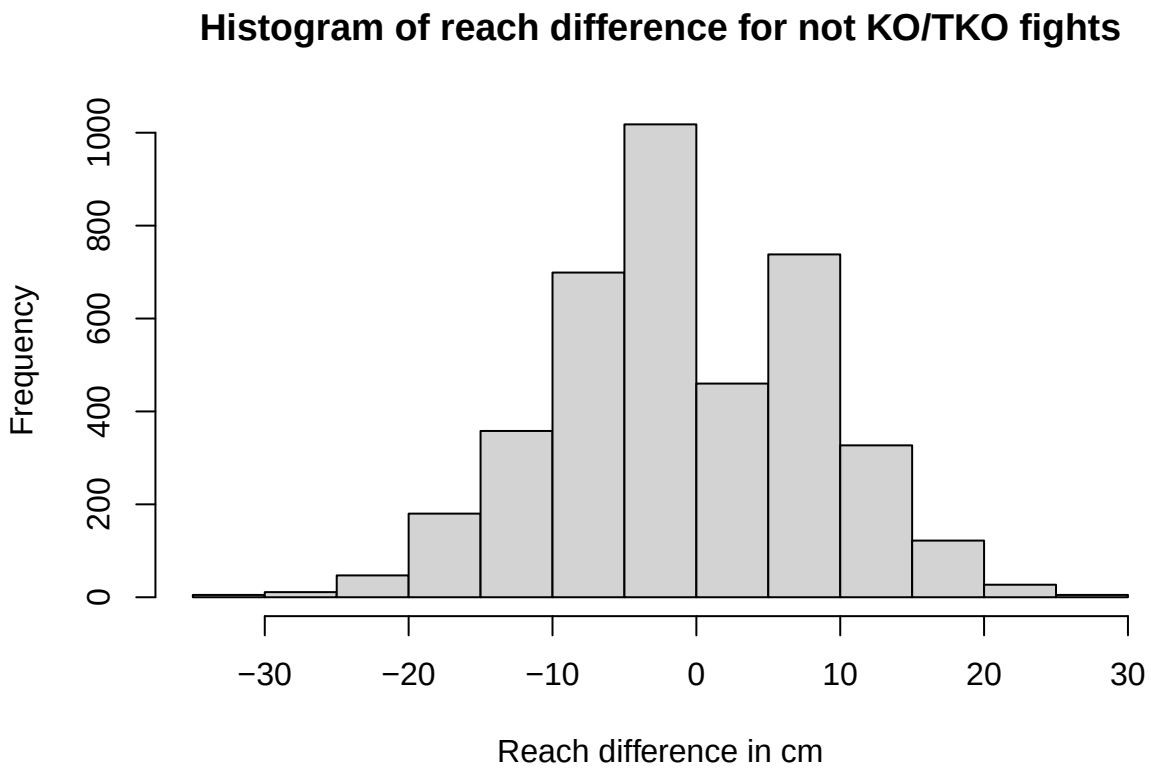
Kako bi mogli provesti test, moramo najprije provjeriti pretpostavke normalnosti i nezavisnosti uzorka. Obzirom da razmatramo dva različita borca, možemo pretpostaviti njihovu nezavisnost. Sljedeći korak je

provjeriti normalnost podataka koju najčešće provjeravamo: histgoramom, qq-plotom te KS-testom (kojim provjeravamo pripadnost podataka distribuciji).

```
par=(mfrow=c(1,2))
hist(ko$razlika,
     main='Histogram of reach difference for KO/TKO fights',
     xlab='Reach difference in cm')
```

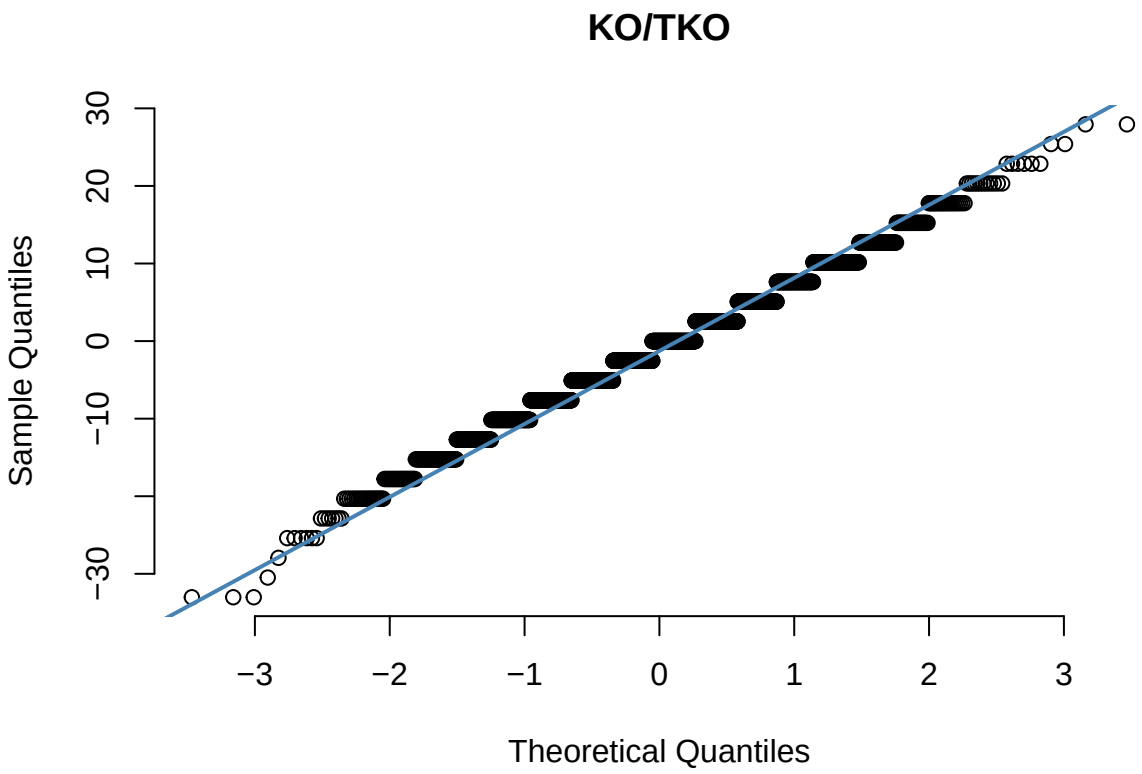



```
hist(not_ko$razlika,
     main='Histogram of reach difference for not KO/TKO fights',
     xlab='Reach difference in cm')
```

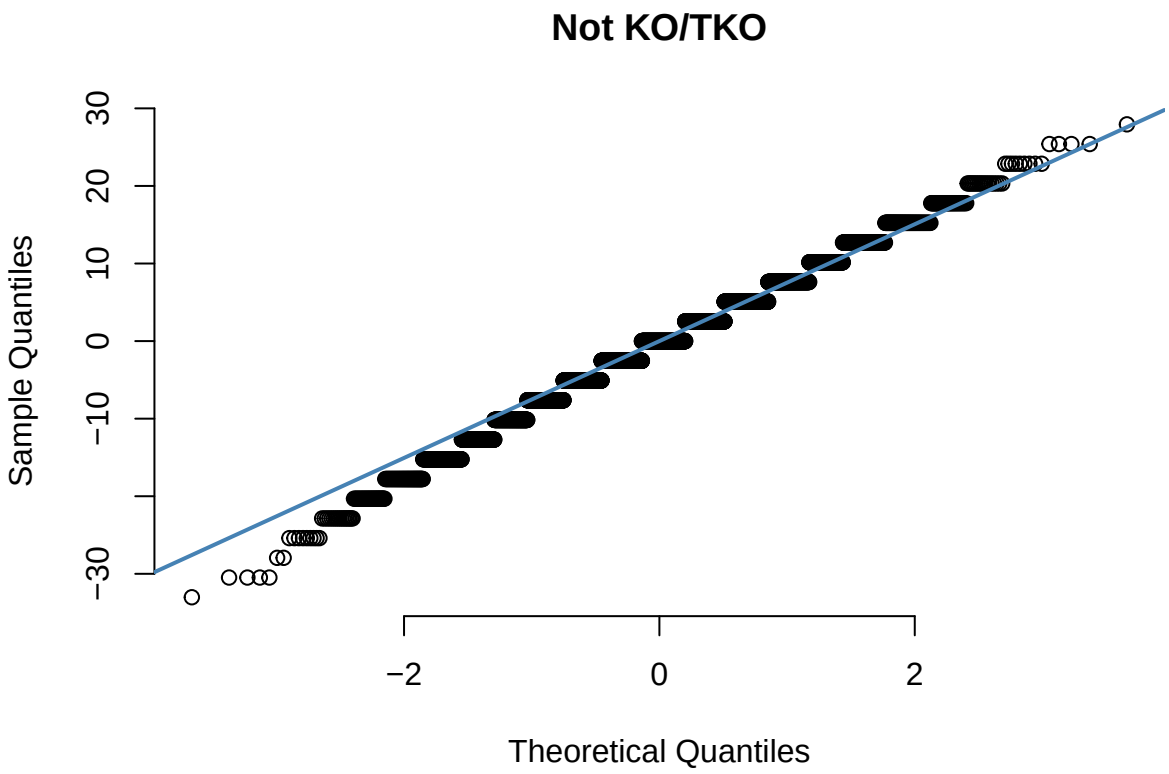


Histogrami upućuju na normalnost podataka. Normalnost možemo još provjeriti i qqplot-ovima ili testom koji ispituje normalnost.

```
par=(mfrow=c(1,2))

qqnorm(ko$razlika, pch = 1, frame = FALSE,main='KO/TKO')
qqline(ko$razlika, col = "steelblue", lwd = 2)
```




```
qqnorm(not_ko$razlika, pch = 1, frame = FALSE,main='Not KO/TKO')
qqline(not_ko$razlika, col = "steelblue", lwd = 2)
```




```
var(ko$razlika)
```

```
## [1] 74.45139
```

```
var(not_ko$razlika)
```

```
## [1] 69.40993
```

Test o jednakosti varijanci

Ako imamo dva nezavisna slučajna uzorka $X_1^1, X_1^2, \dots X_1^{n_1}$ i $X_2^1, X_2^2, \dots, X_2^{n_2}$ koji dolaze iz normalnih distribucija s varijancama σ_1^2 i σ_2^2, tada slučajna varijabla

$$F = \frac{S_{X_1}^2/\sigma_1^2}{S_{X_2}^2/\sigma_2^2}$$

ima Fisherovu distribuciju s $(n_1 - 1, n_2 - 1)$ stupnjeva slobode, pri čemu vrijedi:

$$S_{X_1}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_1^i - \bar{X}_1)^2, \quad S_{X_2}^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_2^i - \bar{X}_2)^2.$$

Hipoteze testa jednakosti varijanci glase:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 < \sigma_2^2 \quad , \quad \sigma_1^2 > \sigma_2^2 \quad , \quad \sigma_1^2 \neq \sigma_2^2$$

```
var.test(ko$razlika, not_ko$razlika)
```

```
##
##  F test to compare two variances
##
## data:  ko$razlika and not_ko$razlika
## F = 1.0726, num df = 1904, denom df = 3996, p-value = 0.07311
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9934681 1.1593477
## sample estimates:
## ratio of variances
##           1.072633
```

p-vrijednost od 0.07311 nam govori da nećemo odbaciti hipotezu H_0 da su varijance naša dva uzorka jednaka.

Provedimo sada t-test uz pretpostavku jednakosti varijanci. Test provodimo uz nivo značajnosti $\alpha = 0.05$.

```
t.test(ko$razlika, not_ko$razlika, alt = "greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  ko$razlika and not_ko$razlika
## t = -2.0636, df = 5900, p-value = 0.9804
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.870258       Inf
## sample estimates:
##  mean of x  mean of y
## -0.9506667 -0.4664398
```

Rezultat

Zbog veće p-vrijednosti (p-value = 0.9804) možemo zadržati H_0 hipotezu o jednakosti prosječnih vrijednosti, odnosno možemo reći da s razlikom u dužini ruku boraca ne možemo očekivati završetak borbe nokautom.

Pitanje 2.

Razlikuje li se trajanje mečeva (u sekundama) između pojedinih kategorija?

Kako bismo provjerili jesu li srednje vrijednosti 3 ili više populacija jednake koristimo analizu varijance.

Na početku je potrebno učitati spojeni skup podataka i izračunati ukupno vrijeme trajanja borbi u novom stupcu ***duration***.

```
data = read.csv(file = "combined.csv")

library(lubridate)
```

```
## Loading required package: timechange
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
data$last_round_time=as.numeric(as.period(ms(data$last_round_time), unit = "sec"))

data$duration = (data$last_round-1)*5*60+data$last_round_time
```

Pretpostavke

Pretpostavke ANOVA-e su:

- normalna razdioba podataka
- homoskedastičnost populacija
- nezavisnost podataka u uzorcima.

ANOVA je robusna na blaga odstupanja od pretpostavke normalnosti i homoskedastičnosti pod uvjetom da su veličine uzoraka podjednake.

U slijedećem isječku koda ćemo provjeriti jesu li veličine uzoraka približno jednake.

```
nrow(data[data$weight_class_Flyweight==1,])
```

```
## [1] 226
```

```
nrow(data[data$weight_class_Bantamweight==1,])
```

```
## [1] 462
```

```
nrow(data[data$weight_class_Featherweight==1,])
```

```
## [1] 539
```

```
nrow(data[data$weight_class_Lightweight==1,])
```

```
## [1] 1072
```

```
nrow(data[data$weight_class_Welterweight==1,])
```

```
## [1] 1066
```

```
nrow(data[data$weight_class_Middleweight==1,])
```

```
## [1] 803
```

```
nrow(data[data$weight_class_LightHeavyweight==1,])
```

```
## [1] 559
```

```
nrow(data[data$weight_class_Heavyweight==1,])
```

```
## [1] 573
```

```
nrow(data[data$weight_class_OpenWeight==1,])
```

```
## [1] 86
```

```
nrow(data[data$weight_class_WomenFlyweight==1,])
```

```
## [1] 110
```

```
nrow(data[data$weight_class_WomenBantamweight==1,])
```

```
## [1] 149
```

```
nrow(data[data$weight_class_WomenFeatherweight==1,])
```

```
## [1] 16
```

```
nrow(data[data$weight_class_WomenStrawweight==1,])
```

```
## [1] 190
```

Vidimo da veličine uzoraka nisu jednake.

Zatim ćemo provjeriti dolaze li podaci iz normalnih distribucija Lillieforsovom inačicom Kolmogorov-Smirnovljevog testa.

```
require(nortest)
```

```
## Loading required package: nortest
```

```
lillie.test(data$duration)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$duration
## D = 0.24115, p-value < 2.2e-16
```

```
lillie.test(data$duration[data$weight_class_Flyweight==1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$duration[data$weight_class_Flyweight == 1]
## D = 0.30543, p-value < 2.2e-16
```

```
lillie.test(data$duration[data$weight_class_Bantamweight==1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$duration[data$weight_class_Bantamweight == 1]
## D = 0.27288, p-value < 2.2e-16
```

```
lillie.test(data$duration[data$weight_class_Featherweight==1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$duration[data$weight_class_Featherweight == 1]
## D = 0.2944, p-value < 2.2e-16
```

```
lillie.test(data$duration[data$weight_class_Lightweight==1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$duration[data$weight_class_Lightweight == 1]
## D = 0.27112, p-value < 2.2e-16
```

```
lillie.test(data$duration[data$weight_class_Welterweight==1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$duration[data$weight_class_Welterweight == 1]
## D = 0.24605, p-value < 2.2e-16
```

```
lillie.test(data$duration[data$weight_class_Middleweight==1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$duration[data$weight_class_Middleweight == 1]
## D = 0.20364, p-value < 2.2e-16
```

```
lillie.test(data$duration[data$weight_class_LightHeavyweight==1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$duration[data$weight_class_LightHeavyweight == 1]
## D = 0.20125, p-value < 2.2e-16
```

```
lillie.test(data$duration[data$weight_class_Heavyweight==1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$duration[data$weight_class_Heavyweight == 1]
## D = 0.12936, p-value < 2.2e-16
```

```
lillie.test(data$duration[data$weight_class_WomenStrawweight==1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$duration[data$weight_class_WomenStrawweight == 1]
## D = 0.35527, p-value < 2.2e-16
```

```
lillie.test(data$duration[data$weight_class_WomenFlyweight==1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$duration[data$weight_class_WomenFlyweight == 1]
## D = 0.30419, p-value < 2.2e-16
```

```
lillie.test(data$duration[data$weight_class_WomenBantamweight==1])
```

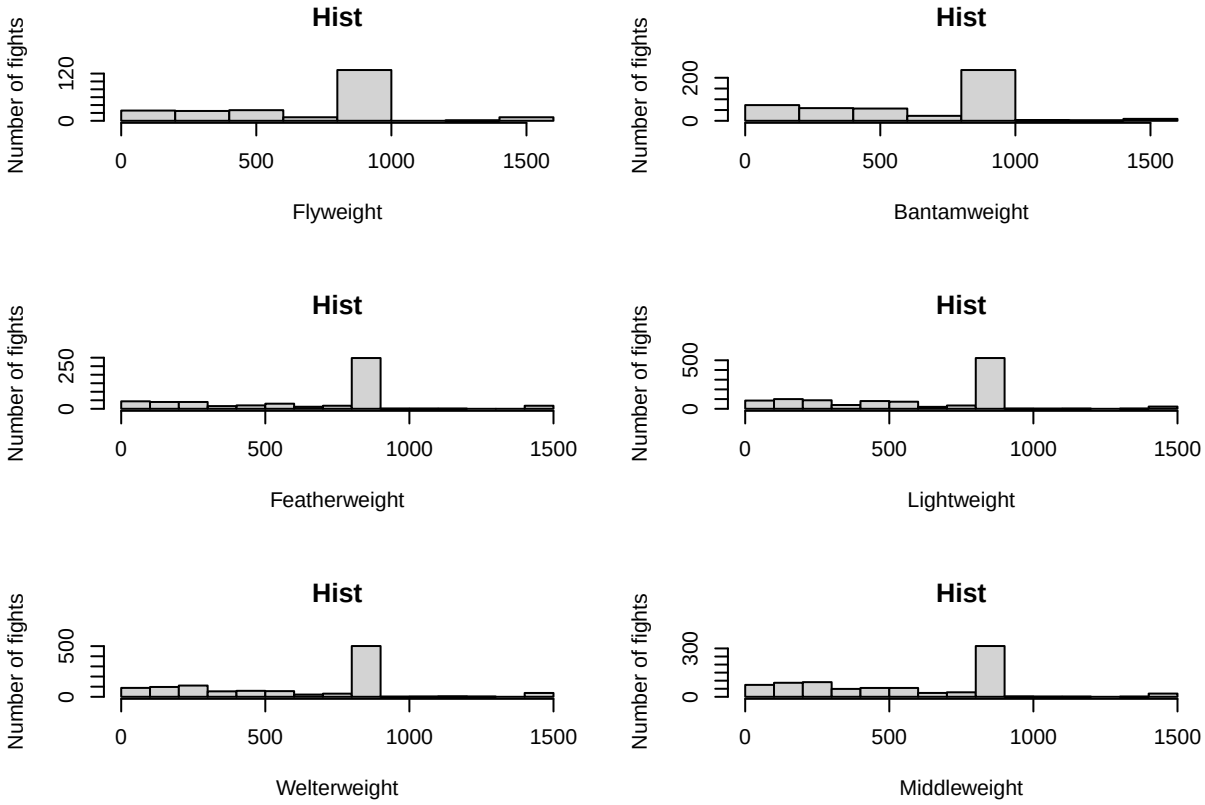
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$duration[data$weight_class_WomenBantamweight == 1]
## D = 0.3205, p-value < 2.2e-16
```

```
lillie.test(data$duration[data$weight_class_WomenFeatherweight==1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$duration[data$weight_class_WomenFeatherweight == 1]
## D = 0.25308, p-value = 0.007279
```

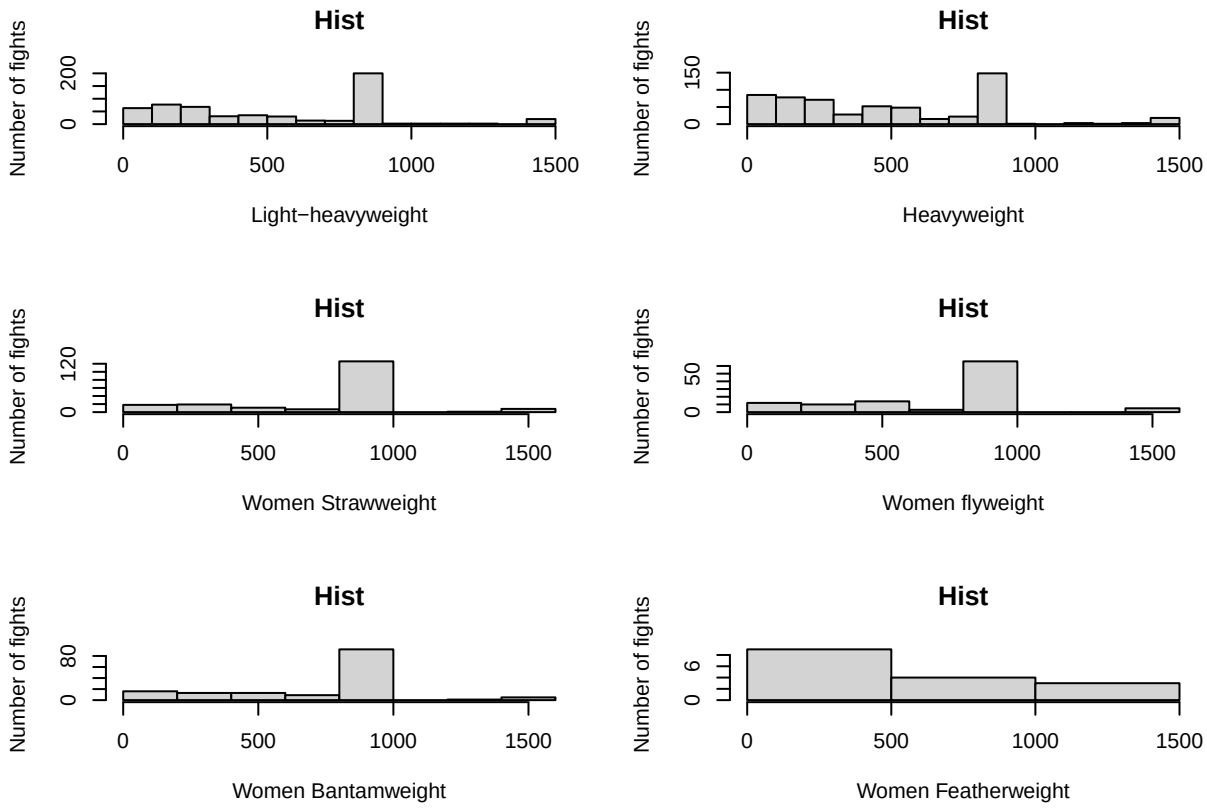
Intuitivnu potvrdu rezultata iz prethodnog testa možemo vidjeti u slijedećim histogramima.

```
par(mfrow=c(3,2))

hist(data$duration[data$weight_class_Flyweight==1],
     main="Hist", xlab="Flyweight", ylab="Number of fights")
hist(data$duration[data$weight_class_Bantamweight==1],
     main="Hist", xlab="Bantamweight", ylab="Number of fights")
hist(data$duration[data$weight_class_Featherweight==1],
     main="Hist", xlab="Featherweight", ylab="Number of fights")
hist(data$duration[data$weight_class_Lightweight==1],
     main="Hist", xlab="Lightweight", ylab="Number of fights")
hist(data$duration[data$weight_class_Welterweight==1],
     main="Hist", xlab="Welterweight", ylab="Number of fights")
hist(data$duration[data$weight_class_Middleweight==1],
     main="Hist", xlab="Middleweight", ylab="Number of fights")
```




```
par(mfrow=c(3,2))
hist(data$duration[data$weight_class_LightHeavyweight==1],
     main="Hist", xlab="Light-heavyweight", ylab="Number of fights")
hist(data$duration[data$weight_class_Heavyweight==1],
     main="Hist", xlab="Heavyweight", ylab="Number of fights")

hist(data$duration[data$weight_class_WomenStrawweight==1],
     main="Hist", xlab="Women Strawweight", ylab="Number of fights")
hist(data$duration[data$weight_class_WomenFlyweight==1],
     main="Hist", xlab="Women flyweight", ylab="Number of fights")
```

```
hist(data$duration[data$weight_class_WomenBantamweight==1],
     main="Hist", xlab="Women Bantamweight", ylab="Number of fights")
hist(data$duration[data$weight_class_WomenFeatherweight==1],
     main="Hist", xlab="Women Featherweight", ylab="Number of fights")
```



Uvodimo značajku weight_class kako bismo elegantnije mogli napisati naredbu za provođenje Bartlettovog testa.

```
data$weight_class<-with(data,ifelse(weight_class_Flyweight == 1, 'Fly',
                        ifelse(weight_class_Bantamweight==1,'Bantam',
                        ifelse(weight_class_Featherweight==1,'Feather',
                        ifelse(weight_class_Lightweight==1,'Light',
                        ifelse(weight_class_Welterweight==1,'Welter',
                        ifelse(weight_class_Middleweight==1,'Middle',
                        ifelse(weight_class_LightHeavyweight==1,'Lheavy',
                        ifelse(weight_class_Heavyweight==1, 'Heavy',
                        ifelse(weight_class_WomenFlyweight==1,'WFly',
                        ifelse(weight_class_WomenBantamweight==1,'WBantam',
                        ifelse(weight_class_WomenFeatherweight==1,'WFeather',
                        ifelse(weight_class_WomenStrawweight==1,'WStraw'
                        ,"Open"))))))))))))
```

Slijedeća pretpostavka koju trebamo provjeriti je homoskedastičnost populacija i nju testiramo Bartlettovim testom.

```
bartlett.test(data$duration ~ data$weight_class)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  data$duration by data$weight_class
## Bartlett's K-squared = 28.438, df = 12, p-value = 0.004771
```

Zaključujemo da populacije nemaju jednake varijance.

```
var((data$duration[data$weight_class_Flyweight == 1]))
```

```
## [1] 111375.4
```

```
var((data$duration[data$weight_class_Bantamweight==1]))
```

```
## [1] 115919.8
```

```
var((data$duration[data$weight_class_Featherweight==1]))
```

```
## [1] 121757.5
```

```
var((data$duration[data$weight_class_Lightweight==1]))
```

```
## [1] 118367.7
```

```
var((data$duration[data$weight_class_Welterweight==1]))
```

```
## [1] 131827.2
```

```
var((data$duration[data$weight_class_Middleweight==1]))
```

```
## [1] 125206
```

```
var((data$duration[data$weight_class_LightHeavyweight==1]))
```

```
## [1] 145085.2
```

```
var((data$duration[data$weight_class_Heavyweight==1]))
```

```
## [1] 138060.4
```

```
var((data$duration[data$weight_class_WomenFlyweight==1]))
```

```
## [1] 108238.4
```

```
var((data$duration[data$weight_class_WomenBantamweight==1]))
```

```
## [1] 107707.7
```

```
var((data$duration[data$weight_class_WomenFeatherweight==1]))
```

```
## [1] 286926.1
```

```
var((data$duration[data$weight_class_WomenStrawweight==1]))
```

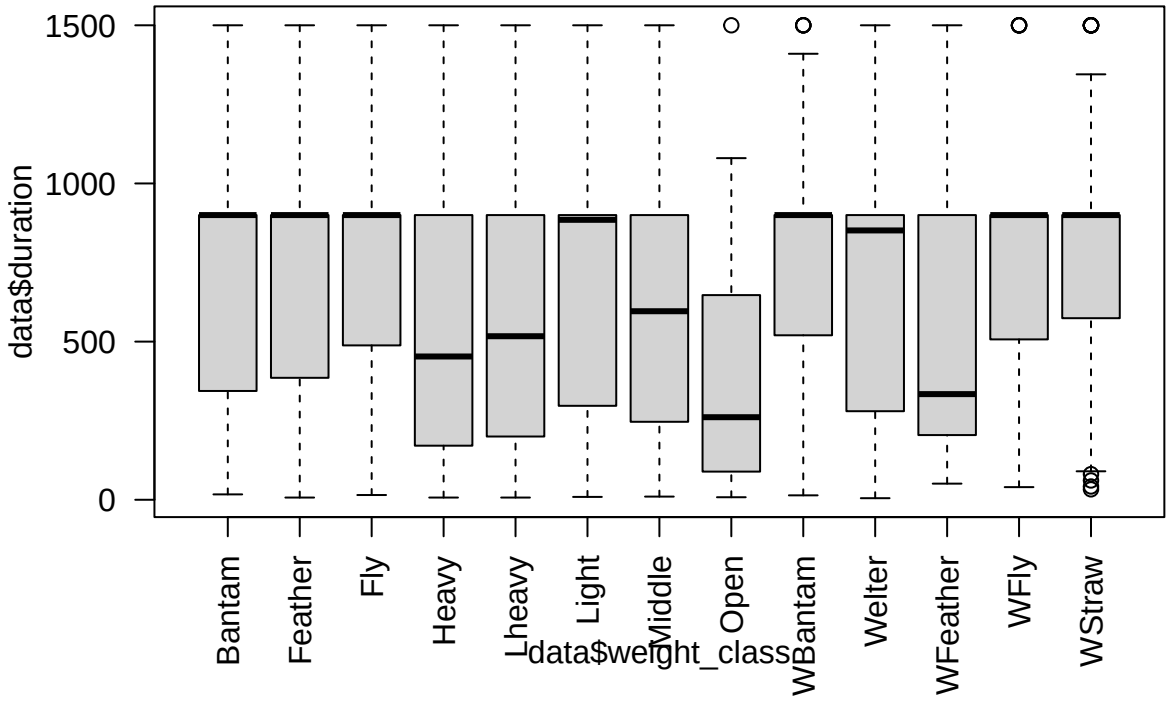
```
## [1] 103843.3
```

```
var((data$duration[data$weight_class_OpenWeight==1]))
```

```
## [1] 64910.04
```

U slijedećim box plotovima možemo vidjeti podatke o trajanju borbi za svaku težinsku kategoriju.

```
# Grafički prikaz podataka
boxplot(data$duration ~ data$weight_class, las=2)
```



Test

U konačnici provodimo ANOVA-u.

Naše hipoteze su:

$H_0 : \mu_1 = \mu_2 = \ldots = \mu 13$

H_1 : Barem jedna od sredina je različita od ostalih.

Test provodimo uz nivo značajnosti od 0.05.

```
# Test
a = aov(data$duration ~ data$weight_class)
summary(a)
```

```
##                      Df    Sum Sq Mean Sq F value Pr(>F)
## data$weight_class   12  31205605 2600467   20.73 <2e-16 ***
## Residuals         5889 738819897  125458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rezultat

Iznimno mala p-vrijednost znači da odbacujemo H_0 u korist H_1, odnosno da prihvaćamo da srednje vrijednosti trajanja borbi nisu jednake u svim kategorijama.

Pitanje 3.

Traju li (u rundama) borbe za titulu duže od ostalih borbi u natjecanju?

Na početku je potrebno razdvojiti podatke na dva skupa: 'borbe za titulu' i 'ostale borbe'.

```
ufc=read.csv("combined.csv")

ufc_title=ufc %>% filter(grepl("Title",Fight_type,ignore.case = TRUE))
ufc_not_title=ufc %>% filter(!grepl("Title",Fight_type,ignore.case = TRUE))
```

```
mean(ufc_title$last_round)
```

```
## [1] 2.983333
```

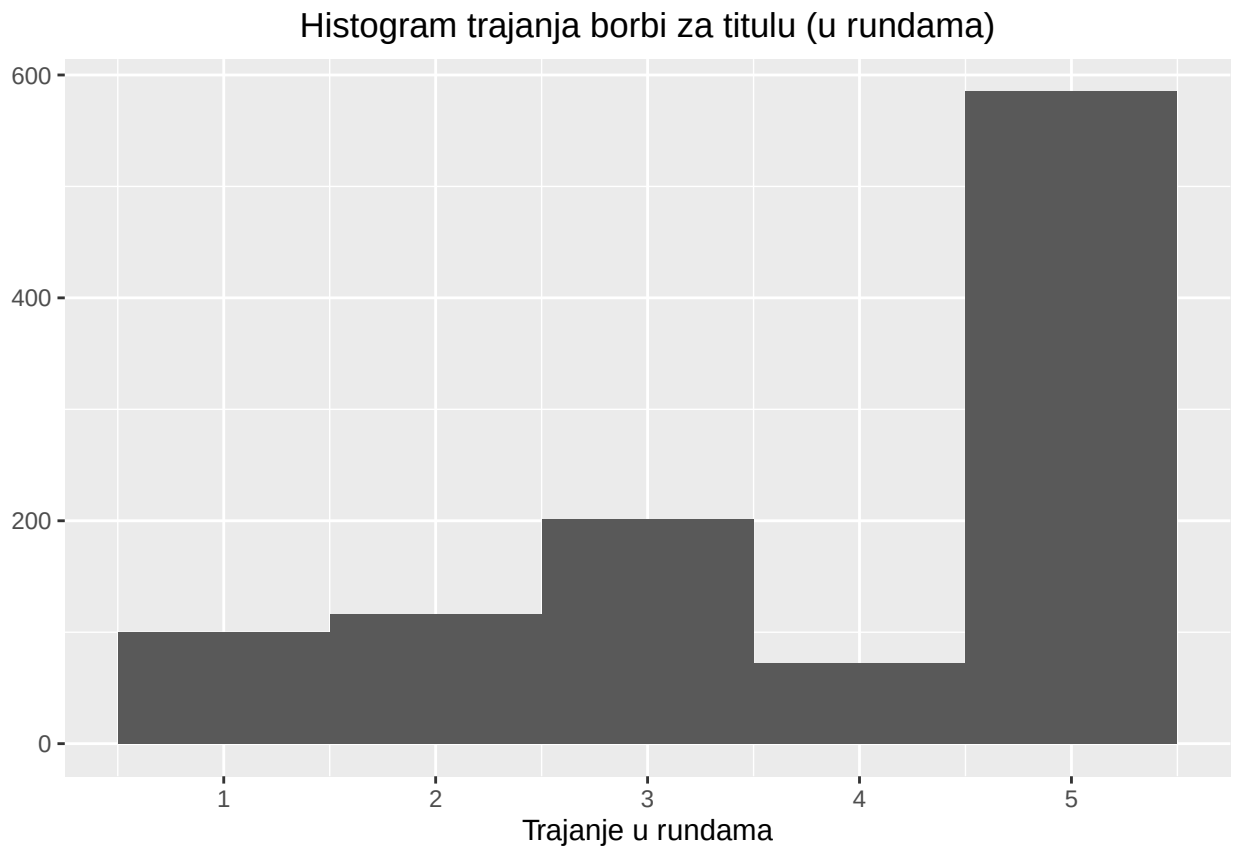
```
mean(ufc_not_title$last_round)
```

```
## [1] 2.27048
```

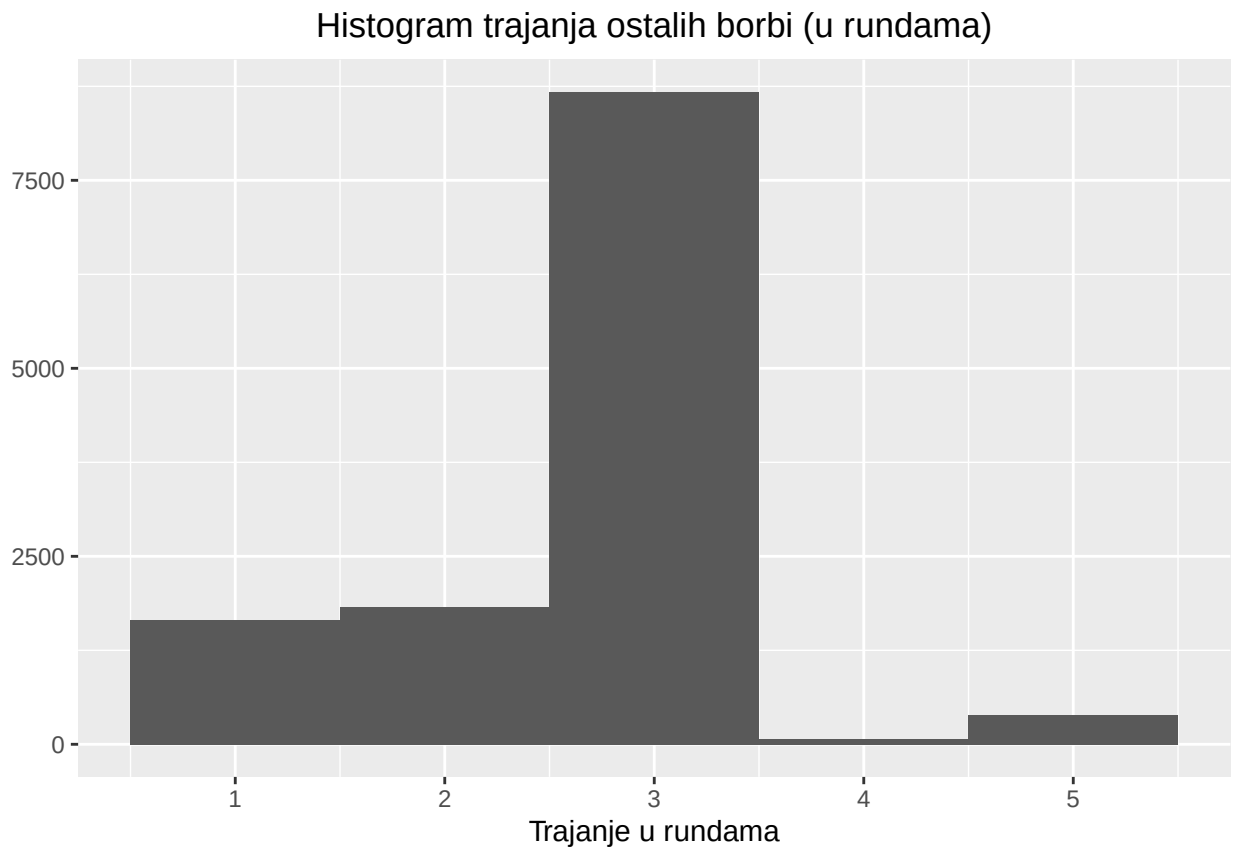
Na prvi pogled uistinu izgleda da su borbe za titulu u prosjeku duže od ostalih borbi. Pomoću statističkog testa možemo vidjeti je li ta razlika značajna.

Prikažimo podatke grafički:

```
df=data.frame(ufc_title$last_round)
library(ggplot2)
p=ggplot(df,aes(ufc_title$last_round,ufc_title$last_round))+geom_bar(stat="identity",width=1)+ggtitle("
p + theme(
plot.title = element_text(hjust = 0.5),
axis.title.y = element_blank()
)
```




```
df=data.frame(ufc_not_title$last_round)
p=ggplot(df,aes(ufc_not_title$last_round,ufc_not_title$last_round))+geom_bar(stat="identity",width=1)+g
p + theme(
plot.title = element_text(hjust = 0.5),
axis.title.y = element_blank()
)
```



Pošto se radi o diskretnim podacima, oni ne mogu biti normalno raspoređeni. U tom slučaju koriste se neparametarski testovi.

```
nrow(ufc_title)
```

```
## [1] 360
```

```
nrow(ufc_not_title)
```

```
## [1] 5542
```

Pošto je broj uzoraka jedne populacije veći od broja uzoraka druge, koristimo Mann-Whitney-Wilcoxonov test i možemo pretpostaviti nezavisnost uzoraka.

Hipoteze su:

$H_0 : \tilde{\mu_1} = \tilde{\mu_2}$

$H_1 : \tilde{\mu_1} > \tilde{\mu_2}$

Test provodimo uz nivo značajnosti od 0.05.

```
wilcox.test(ufc_title$last_round,ufc_not_title$last_round,alternative = "greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  ufc_title$last_round and ufc_not_title$last_round
## W = 1215163, p-value = 1.986e-14
## alternative hypothesis: true location shift is greater than 0
```

Rezultat

Odbacujemo H_0 u korist H_1: medijan trajanja(u rundama) borba za titulu veći je od medijana trajanja ostalih borba.

Pitanje 4.

Mogu li dostupne značajke predvidjeti pobjednika?

```
ufcBorbe <- read.csv("UFC.csv")
# dim(ufcBorbe)       # broj redaka, broj stupaca (broj primjera, broj varijabli)
# nrow(ufcBorbe)      # broj redaka
# ncol(ufcBorbe)      # broj stupaca -> što daje length?
colNames <- names(ufcBorbe)     # imena stupaca
# print(colNames)
```

Iz učitanih podataka, može se vidjeti da će stupac ***Winner*** sadržavati vrijednosti koje želimo predvidjeti pomoću ostalih značajki. Također je potrebno zamijeniti vrijednosti stupca ***title_bout*** i ***Winner***, true i false, s vrijednostima 1 i 0 respektivno.

Isto tako, potrebno je odrediti koje varijable su međusobno zavisne jer njihova zavisnost može dati netočne i neprecizne rezultate. Da bi odredili koje su varijable međusobno zavisne, za provjeru koreliranosti varijabli, koristili smo Pearsonov koeficijent korelacije. Pearsonov koeficijent korelacije r računa koliko su dvije varijable jako povezane i u kojem smjeru. Njegova vrijedost je u intervalu <-1, 1> te što je bliže 1 ili -1 to je zavisnost varijabli veća, takve varijable želimo izbaciti.

```
newData <- data.frame(matrix(nrow = nrow(ufcBorbe)))
newData[colNames[1]] <- ufcBorbe[, 1]   # izlazna varijabla - stupac Winner
newData[colNames[2]] <- ufcBorbe[, 2]   # title_bout

#Promjena vrijednosti značajki Winnier i title_bout
newData$Winner[newData$Winner=="Blue"]<-1
newData$Winner[newData$Winner=="Red"]<-0
newData$Winner<-as.numeric(newData$Winner)
newData$title_bout[newData$title_bout==TRUE]<-1
newData$title_bout[newData$title_bout==FALSE]<-0
newData$title_bout<-as.numeric(newData$title_bout)


prevelikCoef <- FALSE
for(i in 3:160) {
    x <- as.numeric(ufcBorbe[, i])  # novi stupac koji pokusavamo dodati ako nije liearno zavisan o pri
    for(j in 2:ncol(newData)) {
        if(!prevelikCoef) {
            y <- as.numeric(newData[, j])
```

```
            suppressWarnings(coef1 <- cor(x, y, method = "pearson"))
            if((!is.na(coef1)) && ((coef1 < -0.6) || (coef1 > 0.6))) {
                # print(paste0("Usao za ", colNames[i]))
                prevelikCoef <- TRUE
            }
        } else {
            break
        }
    }
    if(!prevelikCoef) {
        newData[colNames[i]] <- ufcBorbe[, i]
    }
    prevelikCoef <- FALSE
}

newData <- newData[ , ! names(newData) %in% c("matrix.nrow...nrow.ufcBorbe..")]

# write.table(newData, file = "tablica.txt", sep = ",")
```

Logistička regresija

Nakon što je određeno koje značajke ćemo uzeti, potrebno je istrenirati model na određenom skupu podataka a onda testirati njegovu točnost na drugim podacima. U tu svrhu, smo ulazni skup podataka podijelili na skup za treniranje (70 posto ulaznih podataka) i skup za testiranje (preostalih 30 posto).

Pozivom ugrađene funkcije za treniranje generičkih linearnih modela, ***glm***, kojoj smo proslijedili sve značajke koje ćemo koristit za predviđanje izlazne varijable.

```
require(caret)
```

```
## Loading required package: caret
```

```
#make this example reproducible
set.seed(520)

#Use 70% of dataset as training set and remaining 30% as testing set
sample1 <- sample(c(TRUE, FALSE), nrow(newData), replace = TRUE, prob = c(0.7, 0.3))
trainSet <- newData[sample1, ]
testSet <- newData[!sample1, ]

logreg.mdl = glm(Winner ~ title_bout + B_avg_KD + B_avg_opp_KD + B_avg_SIG_STR_pct + B_avg_opp_SIG_STR_
                 data = trainSet, family = binomial())

# summary(logreg.mdl)

# Rsq = 1 - logreg.mdl$deviance/logreg.mdl$null.deviance
# Rsq
```

Predikcija

Sada naučeni model možemo koristiti za predviđanje još neviđenih podataka, skup podataka za testiranje. Kao rezultat toga ćemo dobiti matricu konfuzije, oblika

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	TN	FP
$Y = 1$	FN	TP

Dakle pokazuje za koliko primjera, koji su trebali imati izlaz 1, je model vratio 1 (true positive, TP), za koje je vratio 0 (false negative, FN). Isto vrijedi i za promjere s izlazom 0, ako je za njih model vratio isto 0 (true negative, TN) ili je vratio 1 (false positive, FP).

Na temelju tih vrijednosti možemo izračunati sljedeće vrijednosti: - točnost (eng. accuracy): $\frac{TP+TN}{TP+FP+TN+FN}$ - preciznost (eng. precision): $\frac{TP}{TP+FP}$ - odziv (eng. recall): $\frac{TP}{TP+FN}$ - specifičnost (eng. specificity): $\frac{TN}{TN+FP}$

Točnost govori koliko je od svih primjera točno klasificiranih (dakle da je očekivani izlaz bio 0 ili 1 i da je za njih model isto vratio 0 ili 1 respektativno). Preciznost govori koliko je od svih primjera koje je model klasificirao s izlazom 1, a da je njihov očekivani izlaz bio isto 1. Odziv pokazuje koliko je primjera od svih koji su trebali biti klasificirani kao 1, model stvarno i klasificirao kao 1. Na kraju, specifičnost pokazuje udio primjera koji su točno klasificirani kao 0

```
# predict(logreg.mdl, testSet)


yHat <- logreg.mdl$fitted.values > 0.5
tab <- table(trainSet$Winner, yHat)

tab
```

```
##    yHat
##     FALSE TRUE
##   0  2514  268
##   1   962  380
```

```
accuracy = sum(diag(tab)) / sum(tab)
precision = tab[2,2] / sum(tab[,2])
recall = tab[2,2] / sum(tab[2,])
specificity = tab[1,1] / sum(tab[,1])

accuracy
```

```
## [1] 0.7017459
```

```
precision
```

```
## [1] 0.5864198
```

```
recall
```

```
## [1] 0.2831595
```

```
specificity
```

```
## [1] 0.7232451
```

Test omjera izglednosti (likelihood ratio test)

Moguće je odabrati manji broj značajki, tako da se smanji prihvaćeni interval Pearsonovog koeficijenta korelacije. Dobiveni rezultat bi tada usporedili s prvim modelom i vidjeli koji bolje predviđa. Ovaj pristup se naziva test omjera izglednosti

```
logreg.mdl.2 = glm(Winner ~ title_bout + B_avg_KD + B_avg_opp_KD + B_avg_SIG_STR_pct + B_avg_opp_SIG_STR
                   data = trainSet, family=binomial())

yHat <- logreg.mdl.2$fitted.values > 0.5
tab <- table(trainSet$Winner, yHat)

tab
```

```
##    yHat
##     FALSE TRUE
##   0  2557  225
##   1  1082  260
```

```
accuracy = sum(diag(tab)) / sum(tab)
precision = tab[2,2] / sum(tab[,2])
recall = tab[2,2] / sum(tab[2,])
specificity = tab[1,1] / sum(tab[,1])

accuracy
```

```
## [1] 0.6830747
```

```
precision
```

```
## [1] 0.5360825
```

```
recall
```

```
## [1] 0.1937407
```

```
specificity
```

```
## [1] 0.7026656
```

```
anova(logreg.mdl, logreg.mdl.2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Winner ~ title_bout + B_avg_KD + B_avg_opp_KD + B_avg_SIG_STR_pct +
##     B_avg_opp_SIG_STR_pct + B_avg_TD_pct + B_avg_opp_TD_pct +
##     B_avg_SUB_ATT + B_avg_opp_SUB_ATT + B_avg_REV + B_avg_opp_REV +
##     B_avg_SIG_STR_att + B_avg_opp_TOTAL_STR_landed + B_avg_TD_att +
##     B_avg_opp_TD_att + B_avg_LEG_att + B_avg_opp_LEG_att + B_avg_CLINCH_att +
##     B_avg_opp_CLINCH_att + B_avg_GROUND_att + B_avg_opp_GROUND_att +
```

```
##     B_avg_opp_CTRL_time.seconds. + B_total_rounds_fought + B_total_title_bouts +
##     B_current_win_streak + B_current_lose_streak + B_draw + B_win_by_Decision_Majority +
##     B_win_by_Decision_Split + B_win_by_Submission + B_win_by_TKO_Doctor_Stoppage +
##     B_Height_cms + R_avg_KD + R_avg_opp_KD + R_avg_SIG_STR_pct +
##     R_avg_opp_SIG_STR_pct + R_avg_TD_pct + R_avg_opp_TD_pct +
##     R_avg_SUB_ATT + R_avg_opp_SUB_ATT + R_avg_REV + R_avg_opp_REV +
##     R_avg_SIG_STR_att + R_avg_opp_TOTAL_STR_landed + R_avg_TD_att +
##     R_avg_opp_TD_att + R_avg_LEG_att + R_avg_opp_LEG_att + R_avg_CLINCH_att +
##     R_avg_opp_CLINCH_att + R_avg_GROUND_att + R_avg_opp_GROUND_att +
##     R_total_rounds_fought + R_total_title_bouts + R_current_win_streak +
##     R_current_lose_streak + R_draw + R_win_by_Decision_Majority +
##     R_win_by_Decision_Split + R_win_by_Submission + R_win_by_TKO_Doctor_Stoppage +
##     B_age + R_age + weight_class_Bantamweight + weight_class_CatchWeight +
##     weight_class_Featherweight + weight_class_Flyweight + weight_class_Heavyweight +
##     weight_class_LightHeavyweight + weight_class_Lightweight +
##     weight_class_Middleweight + weight_class_OpenWeight + weight_class_Welterweight +
##     weight_class_WomenBantamweight + weight_class_WomenFeatherweight +
##     weight_class_WomenFlyweight + weight_class_WomenStrawweight +
##     B_Stance_Open.Stance + B_Stance_Orthodox + B_Stance_Sideways +
##     B_Stance_Switch + R_Stance_Open.Stance + R_Stance_Orthodox +
##     R_Stance_Sideways + R_Stance_Switch
## Model 2: Winner ~ title_bout + B_avg_KD + B_avg_opp_KD + B_avg_SIG_STR_pct +
##     B_avg_opp_SIG_STR_pct + B_avg_TD_pct + B_avg_opp_TD_pct +
##     B_avg_SUB_ATT + B_avg_opp_SUB_ATT + B_avg_REV + B_avg_opp_REV +
##     B_avg_SIG_STR_att + B_avg_TD_att + B_avg_opp_TD_att + B_avg_GROUND_att +
##     B_avg_opp_GROUND_att + B_total_rounds_fought + B_current_lose_streak +
##     B_draw + B_win_by_Decision_Majority + B_win_by_TKO_Doctor_Stoppage +
##     B_Height_cms + R_avg_KD + R_avg_opp_KD + R_avg_SIG_STR_pct +
##     R_avg_opp_SIG_STR_pct + R_avg_TD_pct + R_avg_opp_TD_pct +
##     R_avg_SUB_ATT + R_avg_opp_SUB_ATT + R_avg_REV + R_avg_opp_REV +
##     R_avg_SIG_STR_att + R_avg_TD_att + R_avg_opp_TD_att + R_avg_CLINCH_att +
##     R_avg_GROUND_att + R_avg_opp_GROUND_att + R_total_title_bouts +
##     R_current_lose_streak + R_draw + R_win_by_Decision_Majority +
##     R_win_by_TKO_Doctor_Stoppage + B_age + weight_class_Bantamweight +
##     weight_class_CatchWeight + weight_class_Featherweight + weight_class_Flyweight +
##     weight_class_LightHeavyweight + weight_class_Lightweight +
##     weight_class_Middleweight + weight_class_OpenWeight + weight_class_Welterweight +
##     weight_class_WomenBantamweight + weight_class_WomenFeatherweight +
##     weight_class_WomenFlyweight + weight_class_WomenStrawweight +
##     B_Stance_Open.Stance + B_Stance_Orthodox + B_Stance_Sideways +
##     B_Stance_Switch + R_Stance_Open.Stance + R_Stance_Orthodox +
##     R_Stance_Sideways + R_Stance_Switch
##   Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
## 1      4041     4664.8
## 2      4060     4800.3 -19  -135.46 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rezultat

Rezultati pokazuju da su se vrijednosti mjera kvalitete smanjile, pala je točnost i preciznost modela. S obzirom na te rezultate, prijašnji model ***logreg.mdl*** će biti prihvaćen.

Pitanje 5.

Postoji li razlika u udjelu borbi koje završe kao KO/TKO ovisno o tome stoje li borci u istim ili različitim gardovima?

Stvaramo stupac ***stance*** koji ima vrijednost *Same* ako borci stoje u istim gardovima i *different* ako stoje u različitim.

```
data = read.csv(file = "combined.csv")

data$stance = with(data, ifelse(B_Stance_Open.Stance==R_Stance_Open.Stance & B_Stance_Orthodox==R_Stanc
```

```
nrow(data[data$stance=="Same",])
```

```
## [1] 3718
```

```
nrow(data[data$stance=="Same" & data$win_by=="KO/TKO",])
```

```
## [1] 1188
```

```
nrow(data[data$stance=="Different",])
```

```
## [1] 2184
```

```
nrow(data[data$stance=="Different" & data$win_by=="KO/TKO",])
```

```
## [1] 717
```

Test

Kao procjenitelje za proporcije p borbi koje su završile kao KO/TKO koristiti ćemo $\hat{P} = \frac{X}{n}$, gdje X predstavlja broj borbi koje su završile kao KO/TKO od ukupno n borbi.

Prema Centralnom graničnom teoremu znamo da, za dovoljno veliki n, $\hat{P}$ ima aproksimativno normalnu distribuciju s očekivanjem p i varijancom $\frac{pq}{n}$.

Označimo procjenitelj udjela borbi koje završavaju kao KO/TKO i u kojima borci stoje u istim gardovima s $\hat{P}_1$ i procjenitelj udjela borbi koje završavaju kao KO/TKO i u kojima borci stoje u različitim gardovima s $\hat{P}_2$.

Iz stabilnosti normalne slučajne varijable na sume slijedi da $\hat{P}_1 - \hat{P}_2$ također ima aproksimativno normalnu distribuciju s očekivanjem $p_1 - p_2$ i varijancom $\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$.

Za testiranje ćemo koristiti statistiku

$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$.

Naše hipoteze su:

$H_0 : p_1 = p_2$
$H_1 : p_1 \neq p_2$.

Test provodimo uz nivo značajnosti $\alpha = 0.05$ i zbog oblika H_1 koristimo dvostranu alternativu.

```r
res <- prop.test(x = c(1188, 717), n = c(3718, 2184))

res
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(1188, 717) out of c(3718, 2184)
## X-squared = 0.44481, df = 1, p-value = 0.5048
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.03388265  0.01634249
## sample estimates:
##    prop 1    prop 2
## 0.3195266 0.3282967
```

Rezultat

P-vrijednost iznosi 0.5048, te zbog toga uz nivo značajnosti 0.05 ne možemo odbaciti hipotezu H_0. Drugim riječima, prihvaćamo da su udjeli borbi koje završavaju kao KO/TKO jednaki neovisno o razlici u gardovima boraca.

Pitanje 6.

Postoji li razlika u udjelu borbi koje završavaju submissionom ovisno o tome održava li se event u Brazilu ili ne?

Prvo moramo podijeliti podatke po lokaciji.

```r
library(stringr)
data = read.csv(file = "combined.csv")

data_brazil=data[str_detect(data$location, "Brazil"),]
data_not_brazil=data[!str_detect(data$location, "Brazil"),]
```

Pogledajmo ukupan broj borbi u oba skupa i broj borbi koje su završile submissionom.

```r
nrow(data_brazil)
```

```
## [1] 420
```

```r
nrow(data_brazil[data_brazil$win_by=="Submission",])
```

```
## [1] 103
```

```r
nrow(data_not_brazil)
```

```
## [1] 5482
```

```r
nrow(data_not_brazil[data_not_brazil$win_by=="Submission",])
```

```
## [1] 1108
```

Test

Test koji ćemo provesti gotovo je identičan onom u 5. pitanju. Jedina razlika je u hipotezama. Naime, logično bi bilo pretpostaviti da je udio submissiona na eventima u Brazilu veći, pa su naše hipoteze:

$H_0 : p_{Brazil} = p_{Ostalo}$

$H_1 : p_{Brazil} > p_{Ostalo}$.

Test provodimo uz razinu značajnosti 0.05.

```r
res <- prop.test(x = c(103, 1108), n = c(420, 5482), alternative="greater")

res
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(103, 1108) out of c(420, 5482)
## X-squared = 4.1876, df = 1, p-value = 0.02036
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.006176158 1.000000000
## sample estimates:
##    prop 1    prop 2
## 0.2452381 0.2021160
```

Rezultat

p-vrijednost manja od 0.05 znači da odbacujemo H_0 u korist H_1, odnosno prihvaćamo da je udio submissiona veći kod evenata koji se održavaju u Brazilu.