Badanie zawodników NBA pod względem statystyk, które zdobywali na przestrzeni lat

Bartłomiej Ortyl

Informatyka i Ekonometria

Rok 2

Spis treści

1.Wstęp

1.1 Cel pracy i Hipotezy

W moim projekcie zamierzam odpowiedzieć na 6 pytań badawczych dotyczących zawodników NBA oraz wpływu ich wyboru w drafcie na punkty, asysty i rozegrane gry. Zbadam również zawodników prawo lub lewo ręcznych i wpływ mocniejszej ręki na punkty, asysty i rozegrane gry.

HIPOTEZY

H1: Wyższy pick w drafcie przekłada się na wyższe zdobycze punktowe.

H2: Wyższy pick w drafcie przekłada się na wyższe zdobycze asyst.

H3: Wyższy pick w drafcie przekłada się na więcej rozegranych meczów w karierze.

H4: Nie ma znaczenia czy gracz jest lewo lub prawo ręczny, zdobywają tyle samo punktów

H5: Nie ma znaczenia czy gracz jest lewo lub prawo ręczny, zdobywają tyle samo asyst

H6: Nie ma znaczenia czy gracz jest lewo lub prawo ręczny, rozgrywają tyle samo meczy

2.Część teoretyczna

2.1 Metody

W moim projekcie używam:

- analiza wariancji
- test t-studenta
- chi kwadrat test zgodnośći
- regresja liniowa

Test t-Studenta dla prób niezależnych

Test t-Studenta dla prób niezależnych służy do porównywania dwóch grup z populacji mających rozkłady normalne

Założenia testu: homogeniczność wariancji, rozkład normalny zmiennej zależnej, równoliczność grup

Hipoteza zerowa, czyli nie ma różnic pomiędzy porównywanymi grupami i alternatywna, czyli występują różnice pomiędzy porównywanymi grupami

Chi Kwadrat

Test zgodności chi-kwadrat (inaczej zwany testem Pearsona) służy do porównania ze sobą zaobserowanego rozkładu naszej zmiennej z jakimś teoretycznym rozkładem. Jednakże przy testowaniu zgodności rozkładu naszej zmiennej z dobrze znanymi rozkładami teoretycznymi w statystyce : np. normalnym, Poissona zazwyczaj stosuje się inne testy np. test K-S, test Shapiro-Wilka.

Test zgodności chi-kwadrat w praktyce można wykorzystać przynajmniej na dwa sposoby:

-sprawdzenie równoliczności grup
-porównanie występowania obserwacji z ich teoretycznym występowaniem

Analiza warianacji ANOVA

Jedna z najbardziej popularnych i najczęściej stosowanych analiz statystycznych. Dokładniej - analizą wariancji określa się grupę analiz, służących do badania wpływu czynników (zmiennych niezależnych) na zmienną zależną.

Jest to stosunek wariancji, którą obliczyliśmy pomiędzy badanymi grupami a średnią wariancją, którą zaobserwowaliśmy wewnątrz grup (nie mylić z czynnikami między i wewnątrzgrupowymi). Analiza ta jest metodą statystyczną pozwalającą na podział zaobserwowanej zmienności (wariancji) wyników na oddzielne części. Analizowana jest wariancji przypadająca na każdy z analizowanych czynników jak również wariancji błędu.

Regresja Liniowa

Regresja liniowa to najprostszy wariant regresji w statystyce. Zakłada ona, że zależność pomiędzy zmienną objaśnianą a objaśniająca jest zależnością liniową. Tak jak w analizie korelacji, jeżeli jedna wartość wzrasta to druga wzrasta (dodatnia korelacji) lub spada (korelacja ujemna). W regresji liniowej zakłada się, że wzrostowi jednej zmiennej towarzyszy wzrost lub spadek na drugiej zmiennej. Co więcej, nazwa regresji liniowej odnosi się, że funkcja regresji przyjmuje postać funkcji liniowej, czyli **y = bx+a**.

3.Część praktyczna

3.1 Interpretacja Hipotez oraz wnioski z nich płynące

H1:

Wyższy pick w drafcie równa przekłada się na wyższe zdobycze punktowe.

Tę hipotezę przetestowałem za pomocą analizy wariancji. Chodziło tu o porównanie grupy draft 1, 2 i 3 pod względem zdobytych punktów.

Najpierw dokonałem analizy rozkładów za pomocą testu Shapiro-Wilka.

```
#H1
library(car)
shapiro.test(players2$career_PTS[players2$draft_pick=="1st overall"])
shapiro.test(players2$career_PTS[players2$draft_pick=="2nd overall"])
shapiro.test(players2$career_PTS[players2$draft_pick=="3rd overall"])
leveneTest(players2$career_PTS~players2$draft_pick)
model1<- aov(players2$career_PTS~players2$draft_pick)
summary(model1)
```

```
> shapiro.test(players2$career_PTS[players2$draft_pick=="1st overall"]) #ten test trzeba zrobic dla zmiennej zaleznej w kazdej podgrupie

        Shapiro-Wilk normality test

data:  players2$career_PTS[players2$draft_pick == "1st overall"]
W = 0.97687, p-value = 0.2289

> shapiro.test(players2$career_PTS[players2$draft_pick=="2nd overall"])

        Shapiro-Wilk normality test

data:  players2$career_PTS[players2$draft_pick == "2nd overall"]
W = 0.97501, p-value = 0.1746

> shapiro.test(players2$career_PTS[players2$draft_pick=="3rd overall"]) #zalozenie jest spelnione gdy p>0.05 = rozklad jest normalny

        Shapiro-Wilk normality test

data:  players2$career_PTS[players2$draft_pick == "3rd overall"]
W = 0.94601, p-value = 0.004882

> leveneTest(players2$career_PTS~players2$draft_pick)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  1.7784 0.1715
      205
```

```
> model1<- aov(players2$career_PTS~players2$draft_pick)
> summary(model1)
                     Df Sum Sq Mean Sq F value Pr(>F)
players2$draft_pick   2    183   91.54   2.964 0.0538 .
Residuals           205   6331   30.88
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Shapiro-Wilk: punkty mają rozkład normalny dla draftu 1 i 2 (p>0,05), ale nie dla draftu 3 (p=0,005).

Następnie sprawdziłem równość wariancji w podgrupach testem Levene’a.

Wynik testu Levene'a jest nieistotny, czyli wariancje są równe.

Anova : Pr > 0,05, co oznacza że nie ma różnic pomiędzy porównywanymi grupami, co zaprzecza mojej hipotezie.

H2:

Wyższy pick w drafcie równa przekłada się na wyższe zdobycze asyst

Tę hipotezę przetestowałem za pomocą analizy wariancji. Chodziło tu o porównanie grupy draft 1, 2 i 3 pod względem zdobytych asyst.

```
#H2
library(car)
shapiro.test(players2$career_AST[players2$draft_pick=="1st overall"])
shapiro.test(players2$career_AST[players2$draft_pick=="2nd overall"])
shapiro.test(players2$career_AST[players2$draft_pick=="3rd overall"])
leveneTest(players2$career_AST~players2$draft_pick)
model2<- aov(players2$career_AST~players2$draft_pick)
summary(model2)
```

```
> model2<- aov(players2$career_AST~players2$draft_pick)
> shapiro.test(players2$career_AST[players2$draft_pick=="1st overall"]) #ten test trzeba zrob

        Shapiro-Wilk normality test

data:  players2$career_AST[players2$draft_pick == "1st overall"]
W = 0.80852, p-value = 4.782e-08

> shapiro.test(players2$career_AST[players2$draft_pick=="2nd overall"])

        Shapiro-Wilk normality test

data:  players2$career_AST[players2$draft_pick == "2nd overall"]
W = 0.8766, p-value = 5.187e-06

> shapiro.test(players2$career_AST[players2$draft_pick=="3rd overall"])

        Shapiro-Wilk normality test

data:  players2$career_AST[players2$draft_pick == "3rd overall"]
W = 0.87744, p-value = 6.345e-06

> leveneTest(players2$career_AST~players2$draft_pick)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2  0.1867 0.8298
      205
```

```
> summary(model2)
                     Df Sum Sq Mean Sq F value Pr(>F)
players2$draft_pick   2    2.3   1.135   0.281  0.755
Residuals           205  828.4   4.041
>
```

Shapiro-Wilk : Dla żadnego draftu rozkład nie jest normalny. Wszędzie p < 0,05.

Wnioski: We wszystkich przypadkach p<0,05, a więc rozkłady nie są normalne.

Następnie sprawdziłem równość wariancji w podgrupach testem Levene'a.

Wynik testu Levene’a jest nieistotny, czyli wariancje są równe.

Anova : Pr > 0,05, co oznacza że nie ma różnic pomiędzy porównywanymi grupami, co zaprzecza mojej hipotezie.

H3:

Wyższy pick w drafcie równa przekłada się na więcej rozegranych meczów w karierze

Tę hipotezę przetestowałem za pomocą analizy wariancji. Chodziło tu o porównanie grupy draft 1, 2 i 3 pod względem rozegranych meczy.

```
#H3
shapiro.test(players2$career_G[players2$draft_pick=="1st overall"])
shapiro.test(players2$career_G[players2$draft_pick=="2nd overall"])
shapiro.test(players2$career_G[players2$draft_pick=="3rd overall"])
leveneTest(players2$career_G~players2$draft_pick)
model3<- aov(players2$career_G~players2$draft_pick)
summary(model3)
```

```
> model3<- aov(players2$career_G~players2$draft_pick)
> shapiro.test(players2$career_G[players2$draft_pick=="1st overall"]) #ten test trzeba zrobic

        Shapiro-Wilk normality test

data:  players2$career_G[players2$draft_pick == "1st overall"]
W = 0.98369, p-value = 0.5068

> shapiro.test(players2$career_G[players2$draft_pick=="2nd overall"])

        Shapiro-Wilk normality test

data:  players2$career_G[players2$draft_pick == "2nd overall"]
W = 0.96169, p-value = 0.03107

> shapiro.test(players2$career_G[players2$draft_pick=="3rd overall"])

        Shapiro-Wilk normality test

data:  players2$career_G[players2$draft_pick == "3rd overall"]
W = 0.98296, p-value = 0.469

> leveneTest(players2$career_G~players2$draft_pick)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2   1.316 0.2705
      205
```

Shapiro Wilk: dla 1 i 3 draftu p > 0.05, rozkład jest normalny, dla draftu 2

p < 0.05 więc nie jest to rozkład normalny.

Następnie sprawdziłem równość wariancji w podgrupach testem Levene’a.

Wynik testu Levene’a jest nieistotny, czyli wariancje są równe.

```
> summary(model3)
                     Df    Sum Sq Mean Sq F value Pr(>F)
players2$draft_pick   2     75916   37958   0.338  0.713
Residuals           205 23012922  112258
> |
```

ANOVA: Wynik analizy wariancji jest nieistotny (p=0,713), a więc nie ma różnic w meczach pomiędzy różnymi draftami, co zaprzecza mojej hipotezie.

H4:

Nie ma znaczenia czy gracz jest lewo lub prawo ręczny, zdobywają tyle samo punktów

Do sprawdzenia tej hipotezy użyłem testu t-studenta

Jest to test t-Studenta dla prób niezależnych

```
#H4
shapiro.test(players$career_PTS[players$shoots=="Left"])
shapiro.test(players$career_PTS[players$shoots=="Right"])
var.test(players$career_PTS~players$shoots)
t1<-t.test(players$career_PTS~players$shoots, alternative="greater") |
```

```
> players$shoots<-as.factor(players$shoots)
> table(players$shoots)

      Left Left Right      Right
       283          1       4400
> which(players$shoots
+
+ players<-players[-4132, ]
```

```
> players<-players[-4132, ]
> t1<-t.test(players$career_PTS~players$shoots, alternative="greater")
> shapiro.test(players$career_PTS[players$shoots=="Left"]) #gdy p>0.05 - to zalozenie spelnic

        Shapiro-Wilk normality test

data:  players$career_PTS[players$shoots == "Left"]
W = 0.94272, p-value = 4.872e-09

> shapiro.test(players$career_PTS[players$shoots=="Right"])

        Shapiro-Wilk normality test

data:  players$career_PTS[players$shoots == "Right"]
W = 0.89799, p-value < 2.2e-16

> var.test(players$career_PTS~players$shoots)

        F test to compare two variances

data:  players$career_PTS by players$shoots
F = 1.087, num df = 282, denom df = 4399, p-value = 0.3185
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9225198 1.2975265
sample estimates:
ratio of variances
          1.087012
```

```
> t1

        Welch Two Sample t-test

data:  players$career_PTS by players$shoots
t = 3.9651, df = 316.29, p-value = 4.539e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6969167       Inf
sample estimates:
 mean in group Left mean in group Right
           7.545936            6.352477
```

Okazało się, że jeden gracz jest oburęczny więc szybko go usunąłem z danych, żeby dało się przeprowadzić analizę mojej hipotezy. Oraz zmieniłem character na factor.

Testy Shapiro Wilka mają nieistotne wyniki, więc rozkłady nie są normalne.

Var.test p > 0.05, wskazuje na spełnienie założenia

WNIOSKI: Używając testu t-studenta na dwóch grupach zawodników (lewo/prawo ręczni) okazało się , że gracze lewo ręczni różnią się istotnie od

praworęcznych liczbą punktów. Gracze leworęczni rzucają aż o ponad jedno oczko więcej od graczy prawo ręcznych co zaprzecza mojej hipotezie.

H5:

Nie ma znaczenia czy gracz jest lewo lub prawo ręczny, zdobywają tyle samo asyst

Do sprawdzenia tej hipotezy użyłem testu t-studenta

Jest to test t-Studenta dla prób niezależnych

```
#H5

shapiro.test(players$career_AST[players$shoots=="Left"])
shapiro.test(players$career_AST[players$shoots=="Right"])
var.test(players$career_AST~players$shoots)
t2<-t.test(players$career_AST~players$shoots, alternative="greater")
```

```
> t2<-t.test(players$career_AST~players$shoots, alternative="greater")
> shapiro.test(players$career_AST[players$shoots=="Left"]) #gdy p>0.05 - to zalozenie spelnione

        Shapiro-Wilk normality test

data:  players$career_AST[players$shoots == "Left"]
W = 0.85726, p-value = 1.743e-15

> shapiro.test(players$career_AST[players$shoots=="Right"])

        Shapiro-Wilk normality test

data:  players$career_AST[players$shoots == "Right"]
W = 0.81749, p-value < 2.2e-16

> var.test(players$career_AST~players$shoots)

        F test to compare two variances

data:  players$career_AST by players$shoots
F = 1.3908, num df = 282, denom df = 4399, p-value = 6.012e-05
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.180352 1.660168
sample estimates:
ratio of variances
          1.390818
```

```
> t2

        Welch Two Sample t-test

data:  players$career_AST by players$shoots
t = 3.739, df = 308.64, p-value = 0.0001101
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.2004457       Inf
sample estimates:
 mean in group Left mean in group Right
           1.756890            1.398159
```

Testy shapiro wilka wskazują na to, że rozkłady nie są normalne

Var.test $p < 0.05$, wskazuje na niespełnienie założenia

WNIOSKI: Używając testu t-studenta na dwóch grupach zawodników (lewo/prawo ręczni) okazało się , że gracze lewo ręczni istotnie różnią się od praworęcznych liczbą asyst. Gracze leworęczni zdobywają prawie 0.5 asysty więcej ! od graczy prawo ręcznych co ponownie zaprzecza mojej hipotezie.

H6:

Nie ma znaczenia czy gracz jest lewo lub prawo ręczny, rozgrywają tyle samo meczy (lewa/prawa)

Do sprawdzenia tej hipotezy użyłem testu t-studenta

Jest to test t-Studenta dla prób niezależnych

```
#H6
shapiro.test(players$career_G[players$shoots=="Left"])
shapiro.test(players$career_G[players$shoots=="Right"])
var.test(players$career_G~players$shoots)
t3<-t.test(players$career_G~players$shoots, alternative="greater")
```

```
> t3<-t.test(players$career_G~players$shoots, alternative="greater")
> shapiro.test(players$career_G[players$shoots=="Left"]) #gdy p>0.05 - to zalozenie spelnione

        Shapiro-Wilk normality test

data:  players$career_G[players$shoots == "Left"]
W = 0.88191, p-value = 5.49e-14

> shapiro.test(players$career_G[players$shoots=="Right"])

        Shapiro-Wilk normality test

data:  players$career_G[players$shoots == "Right"]
W = 0.814, p-value < 2.2e-16

> var.test(players$career_G~players$shoots)

        F test to compare two variances

data:  players$career_G by players$shoots
F = 1.1923, num df = 282, denom df = 4399, p-value = 0.03548
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.011879 1.423210
sample estimates:
ratio of variances
          1.192304
```

```
> t3

        Welch Two Sample t-test

data:  players$career_G by players$shoots
t = 4.3982, df = 313.19, p-value = 7.489e-06
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 56.52918      Inf
sample estimates:
 mean in group Left mean in group Right
           359.2191            268.7586
```

Testy shapiro wilka wskazują na to, że rozkłady nie są normalne

Var.test $p < 0.05$, wskazuje na niespełnienie założenia

WNIOSKI: Używając testu t-studenta na dwóch grupach zawodników (lewo/prawo ręczni) okazało się , że gracze leworęczni różnią się istotnie od praworęcznych liczbą rozegranych meczy. Gracze leworęczni rozgrywają średnio prawie 100 więcej meczy w przeciągu swojej kariery ! od graczy prawo ręcznych co po raz kolejny zaprzecza mojej hipotezie.

3.2 Metoda Chi-kwadrat

HIPOTEZA: Graczy wybieranych w drafcie przypisywano do zespołów NBA z różną czestotliwością.

```r
#chi-kwadrat

tabelka<-as.data.frame(table(players$draft_team))
tabelka$p<-rep(1/65, 65)
chisq.test(tabelka$Freq, p=tabelka$p)
```

```
> tabelka<-as.data.frame(table(players$draft_team))
> tabelka$p<-rep(1/65, 65)
> chisq.test(tabelka$Freq, p=tabelka$p)

        Chi-squared test for given probabilities

data:  tabelka$Freq
X-squared = 2952, df = 64, p-value < 2.2e-16
```

players | G3_Ortyl_Bartłomiej_400352.R | tabelka | player

Filter

	Var1	Freq	p
1	Atlanta Hawks	127	0.01538462
2	Baltimore Bullets	76	0.01538462
3	Boston Celtics	172	0.01538462
4	Brooklyn Nets	9	0.01538462
5	Buffalo Braves	20	0.01538462
6	Capital Bullets	3	0.01538462
7	Charlotte Bobcats	19	0.01538462
8	Charlotte Hornets	30	0.01538462
9	Chicago Bulls	157	0.01538462
10	Chicago Packers	6	0.01538462
11	Chicago Stags	18	0.01538462
12	Chicago Zephyrs	5	0.01538462
13	Cincinnati Royals	65	0.01538462
14	Cleveland Cavaliers	105	0.01538462
15	Dallas Mavericks	75	0.01538462
16	Denver Nuggets	77	0.01538462
17	Detroit Pistons	153	0.01538462
18	Fort Wayne Pistons	23	0.01538462
19	Golden State Warriors	114	0.01538462
20	Houston Rockets	90	0.01538462

Showing 1 to 22 of 65 entries, 3 total columns

WNIOSKI: Moje wnioski płynące z tego testu: Wynik jest istotny, co oznacza, że kategorie występują z różnym prawdopodobieństwem, czyli do gracze wybierani w drafcie na przestrzeni lat byli przypisywani z różną częstotliwością co potwierdza moją hipotezę.

3.3 Regresja liniowa

HIPOTEZA: Da się przewidzieć na podstawie ilości rozegranych gier, ile zawodnik zdobędzie zdobywał średnio punktów co gre.

```
#regresja
y=ax+b
points ~ gry
model4<-lm(career_PTS~career_G,data=players) #da sie przewidziec punkty przez gry
summary(model4)
punkty = 3.43 + 0.01*career_G
im wyzsze career_G, tym wiecej punkty
```

```
> model4<-lm(career_PTS~career_G,data=players)
> summary(model4)

Call:
lm(formula = career_PTS ~ career_G, data = players)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6688  -1.9691  -0.5947   1.3875  19.1465

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.4253683  0.0642472   53.31   <2e-16 ***
career_G    0.0109373  0.0001549   70.60   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.298 on 4682 degrees of freedom
Multiple R-squared:  0.5156,    Adjusted R-squared:  0.5155
F-statistic:  4984 on 1 and 4682 DF,  p-value: < 2.2e-16
```

WNIOSKI: Da się przewidzieć punkty przez liczbę rozegranych meczy, co potwierdza moją hipotezę.

Punkty = 3.43 + 0,01*career_G

Im więcej rozegranych gier, tym więcej zawodnik średnio będzie zdobywał punktów co gre.

PODSUMOWANIE HIPOTEZ

Pierwsze sześć hipotez które postawiłem, nie sprawdziły się. Okazało się, że wybór gracza w drafcie od 1 do 3 nie ma żadnego przełożenia na to czy zdobywa on więcej punktów, asyst lub czy rozgrywa on więcej spotkań.

Wbrew moim przewidywaniom dotyczących hipotez od 4 do 6 gracze lewo ręczni są zwyczajnie lepsi w najlepszej lidze koszykarskiej świata. Zdobywają oni więcej punktów oraz asyst niż gracze prawo ręczni, oraz podsumowując ich kariery, gracze lewo ręczni grają więcej meczy.

Ostatnie 2 hipotezy dotyczące kolejno, przydzielania graczy do drużyn i przewidywania średniej zdobyczy punktowej na mecz zależnej od rozegranych spotkań sprawdziły się. Gracze są wybierani z różną czestotliwością do drużyn, a na podstawie ilości rozegranych gier możemy prognozować ile dany zawodnik będzie rzucał średnio punktów na mecz.

4.Bibliografia

„Wielka księga koszykówki” – Bill Simmons

„Advanced NBA Stats for Dummies: How to Understand the New Hoops Math” – Bleacher Report, Ehran Khan

„Wykorzystanie testu t Studenta w praktyce” – Rafał Popiel

„Usługi i pomoc przy statystyce i analizie danych” – Opracowania statystyczne Wrocław

Dane wykorzystane w projekcie: https://www.kaggle.com/