Modele i Wnioskowanie Statystyczne

Projekt – sprawozdanie

Semestr 18L

Igor Markiewicz

Spis treści

1 Opis zadania **2**

2 Opis danych **2**
2.1 Histogramy pól rekordów 3

3 Rozwiązanie zadania **5**
3.1 Metryka 5
3.2 Histogramy metryki 6
3.3 Testy statystyczne 7
3.3.1 Test t Welcha 7
3.3.2 Test normalności danych Shapiro – Wilka 7
3.3.3 Test Manna – Whitneya – Wilcoxona (test sumy rang Wilcoxona) 8

4 Bibliografia **9**

1 Opis zadania

Celem niniejszego projektu jest sprawdzenie w której z grup przewaga gry na własnym boisku ma większe znaczenie, w sensie liczby wygranych. Jako grupy zostały wybrane dwie dywizje z klasyfikacji NBA – atlantycka oraz pacyficzna, a zadanie zostało zrealizowane w języku R.

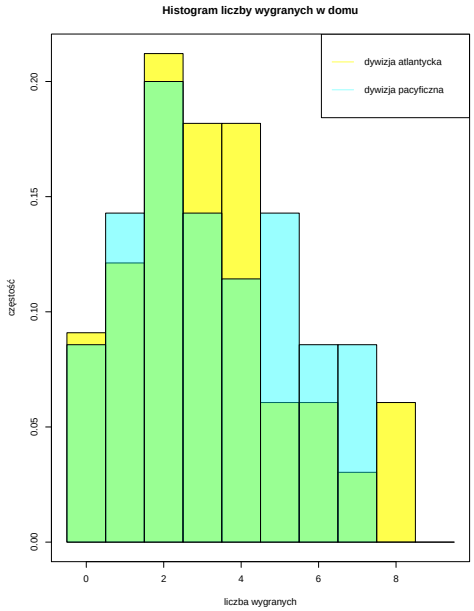
2 Opis danych

Jako dane zostały wybrane informacje ze strony [1] z sezonu 2016/2017 w klasyfikacji regularnej z całego okresu. Każda dywizja posiada pięć zespołów (ich nazwy znajdują się w plikach [8, 9]), a rozgrywki prowadzone są przez okres 7 miesięcy (styczeń – kwiecień i październik – grudzień), dlatego też dane zwierają po ok. 35 rekordów (dla każdej drużyny w danej dywizji z każdego miesiąca), a ich mniejsza liczba wynika z częściowego braku informacji (np: istnieją statystyki gry na własnym boisku w danym miesiącu, ale nie u przeciwników), zdecydowano się na odrzucanie takich rekordów jako niekompletnych. Każdy rekord zawiera cztery wartości :

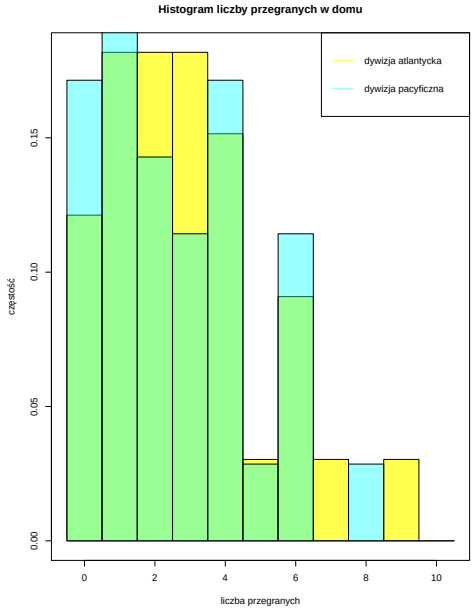
- *WinsHome* – liczba wygranych meczy na własnym boisku
- *LossesHome* – liczba przegranych meczy na własnym boisku
- *WinsRoad* – liczba wygranych meczy u przeciwników
- *LossesRoad* – liczba przegranych meczy u przeciwników

Dane znajdują się w plikach [10], [11], a tagi użyte przy wyszukiwaniu informacji zostały zawarte w pliku [12].

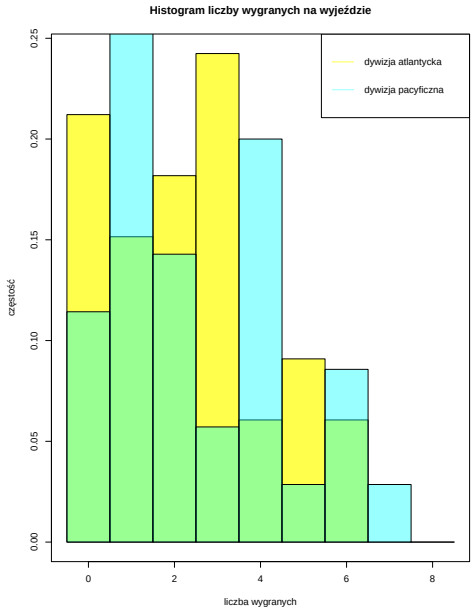
2.1 Histogramy pól rekordów



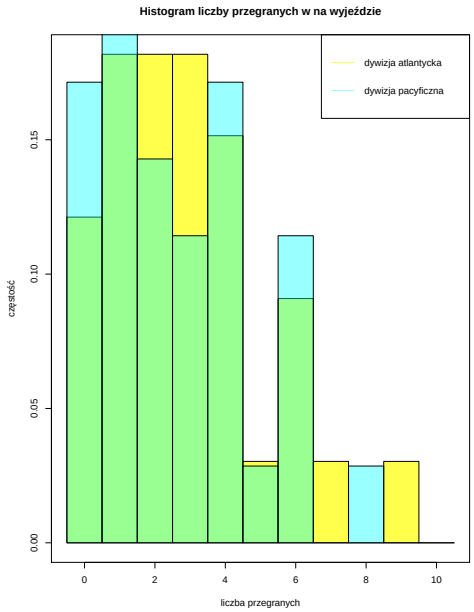
Rys. 1: Liczba wygranych na własnym boisku



Rys. 2: Liczba przegranych na własnym boisku



Rys. 3: Liczba wygranych na wyjeździe



Rys. 4: Liczba przegranych na wyjeździe

3 Rozwiązanie zadania

3.1 Metryka

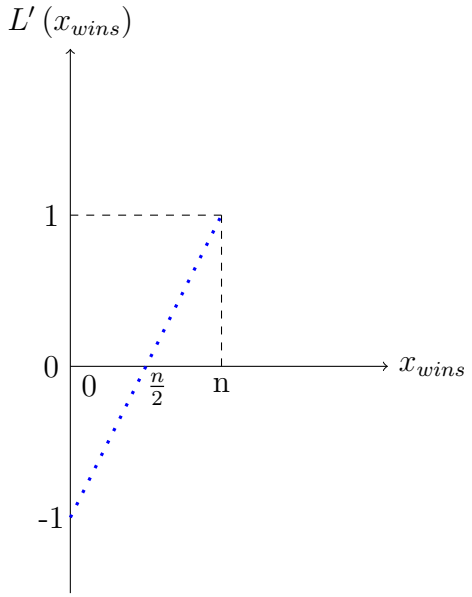
Jako metrykę skaładową zdefiniowano następującą funkcję :

$$L' = \frac{2x_{wins} - n}{n} = \frac{x_{wins} - x_{losses}}{x_{wins} + x_{losses}}$$

Gdzie :

- x_{wins} – liczba wygranych meczy w danym miesiącu, dla danej drużyny i w danym miejscu
- x_{losses} – liczba przegranych meczy w danym miesiącu, dla danej drużyny i w danym miejscu
- $n = x_{wins} + x_{losses}$ – liczba meczy w danym miesiącu, dla danej drużyny i w danym miejscu
- $n \in \mathbb{N}$ oraz $x_{wins}, x_{losses} \in \mathbb{N} \cup \{0\}$
- $L' \in [-1, 1]$

Jako interpetację można wtedy przyjąć że największą (najmniejszą) wartość otrzymujemy wtedy, gdy w danym miesiącu wygraliśmy (przegraliśmy) wszystkie spotkania, natomiast stan pośredni otrzymujemy dla równej liczby wygranych i przegranych meczy.

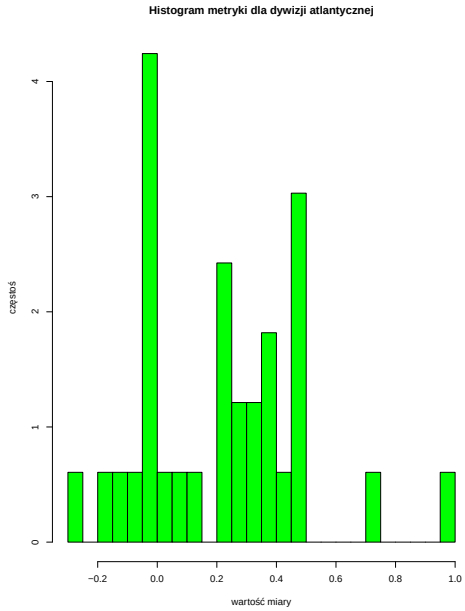


Rys. 5: Wykres $L'(x_{wins})$

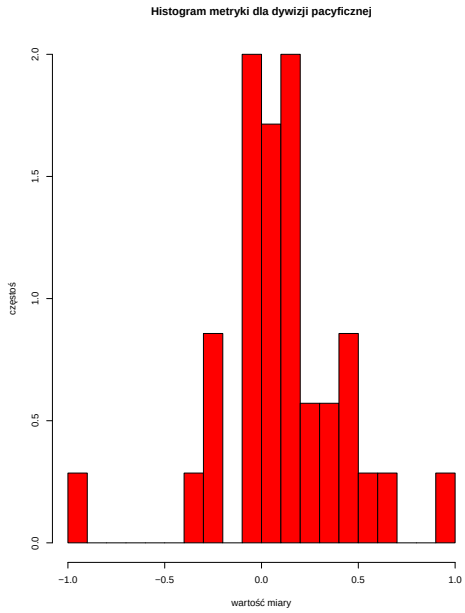
Jako metrykę całkowitą przyjęto połowę różnicy metryk obliczonych na własnym boisku i na wyjeździe (dla poszczególnych drużyn, w poszczególnych miesiącach) :

$$L = \frac{L_{home} - L_{road}}{2}$$
$$L \in [-1, 1]$$

3.2 Histogramy metryki



Rys. 6: Histogram rozkładu metryki dla dywizji atlantyckiej



Rys. 7: Histogram rozkładu metryki dla dywizji pacyficznej

3.3 Testy statystyczne

Skrypt z analizą danych znajduje się w pliku [13].

3.3.1 Test t Welcha

Jako pierwszy test wybrano test *t Welcha* [6, 7] w wersji dwustronnej zakładający rozkład normalny w obu populacjach, oraz brak potrzeby równości wariancj :

$$H_0 - \text{średnie w obu populacjach są równe}$$

$$H_1 - \text{średnie w obu populacjach są różne}$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

$$\nu \approx \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2\nu_1} + \frac{s_2^4}{N_2^2\nu_2}}$$

Gdzie :

- t ma rozkład t – *Studenta* z ν stopniami swobody
- N_1, N_2 – liczebności prób z poszczególnych populacji
- $\overline{X}_1, \overline{X}_2$ – średnie z poszczególnych populacji
- s_1^2, s_2^2 – wariancje z poszczególnych populacji
- $\mu_1 = N_1 - 1, \mu_2 = N_2 - 1$

Zastosowano funkcję *t.test*, z parametrami *alternative="two.sided"* oraz *var.equal=FALSE*. W efekcie otrzymano :

$$p - value = 0,1389$$

co oznacza że dla wartości poziomów istotności $\alpha = 0,05$ oraz $\alpha = 0,1$ nie ma podstaw do odrzucenia hipotezy zerowej.

3.3.2 Test normalności danych Shapiro – Wilka

W celu sprawdzenia normalności danych, zastosowano test Skapiro – Wilka [2, 3] badający czy próbka $x_1, x_2, \ldots, x_n$ pochodzi z populacji o rozkładzie normalnym danej cechy :

$$H_0 - \text{próbka pochodzi z populacji o rozkładzie normalnym}$$

$$H_1 - \text{próbka nie pochodzi z populacji o rozkładzie normalnym}$$

Zastosowano funkcję *shapiro.test*, w efekcie czego otrzymano :

$$p - value = 0.2666$$

dla dywizji atlantyckiej oraz

$$p - value = 0.01739$$

dla dywizji pacyficznej. W związku z tym, w pierwszym przypadku dla poziomów $\alpha = 0,05$ oraz $\alpha = 0,1$ nie ma przeciwskazań do przyjęcia hipotezy zerowej, w drugim zaś możemy odrzucić hipotezę zerową na rzecz alternatywnej.
Test *t Welscha*, ma szansę na poprawne działanie w przypadku, gdy rozkład dla dywizji pacyficznej byłby "blisko" rozkładu noralnego.

3.3.3 Test Manna – Whitneya – Wilcoxona (test sumy rang Wilcoxona)

Na koniec został przeprowadzony test statystyczny sprawdzający czy rozkłady dwóch zbiorów próbek różnią się o stałą wartość μ (przyjętą tutaj jako 0) przy użyciu testu Manna – Whitneya – Wilcoxona [4, 5] (m.in przy założeniu niezależności obserwacji, równej wariancji oraz równości rozkładów) :

$$H_0 - \text{dystrybuanty rozkładów dla dwóch grup są przesunięte o } 0$$

$$H_1 - \text{dystrybuanty rozkładów dla dwóch grup są przesunięte wartość inną niż } 0$$

Zastosowano funkcję *wilcox.test* z parametrem domyślnym *alternative="two.sided"* oraz *paired = FALSE*, w efekcie czego otrzymano :

$$p - value = 0.1761$$

a więc dla poziomów $\alpha = 0,05$ oraz $\alpha = 0,1$ w efekcie nie ma przeciwskazań do przyjęcia hipotezy zerowej.

4 Bibliografia

[1] https://stats.nba.com/

[2] https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

[3] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/shapiro.test.html

[4] https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

[5] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/wilcox.test.html

[6] https://en.wikipedia.org/wiki/Welch%27s_t-test

[7] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/t.test.html

[8] nazwy drużyn dywizji atlantyckiej – *AtlanticDivisionTeams.txt*

[9] nazwy drużyn dywizji pacyficznej – *PacificDivisionTeams.txt*

[10] statystyki dywizji atlantycznej – *AtlanticDivision.csv*

[11] statystyki dywizji pacyficznej – *PacificDivision.csv*

[12] tagi użyte przy poszukiwaniu informacji – *Tags.txt*

[13] skrypt – *script.R*