Autorzy:
Bartłomiej Jaroch
Jakub Gucik

Projekt - Statystyka dla akustyków

Badanie na temat najlepszych zawodników NBA w latach 1991 – 2021 na podstawie nagród MVP

1. Wstęp:

Celem projektu końcowego z przedmiotu było przeprowadzenie badania na powyższy temat i odpowiedzenie na wybrane pytania badawcze w oparciu o dobrane bazy danych. Badanie przeprowadzone przez nasz zespół koncentrowało się na najlepszych zawodnikach ligi NBA w latach 1991 – 2021 wybranych w drodze plebiscytu nagród MVP. Dane do analizy pozyskano ze strony kaggle.com, funkcjonująca jako użyteczna baza wielu danych. Link do wykorzystanych rekordów znajduje się w bibliografii na końcu sprawozdania.

Operowano na danych dotyczących:

- wieku zawodników,
- statystyk skuteczności dotyczących wykonywania rzutów osobistych (linia 4 metrów),
- zapisach średniego spędzonego czasu zawodników na boisku w trakcie meczu,
- zapisach rozegranych meczy na sezon (maksymalnie 82 mecze w sezonie),
- zestawień pozycji, gdzie: C - środkowy; PF – silny skrzydłowy; SF – niski skrzydłowy; SG - rzucający obrońca; PG - rozgrywający.

Wyznaczono również podstawowe cele analizy:

- weryfikacja czy wiek zawodników różnicuje średni czas pobytu na boisku,
- estymowanie skuteczności graczy MVP,
- sprawdzenie, zależności pozycji graczy od ich skuteczności,
- porównanie średnich meczy na sezon klubu Los Angeles Lakers względem pozostałych klubów.

Przeprowadzono, więc analizę struktury (rozkładu zbiorowości względem wybranych cech) oraz współzależności (związków pomiędzy cechami), przy czym warto zauważyć, iż operowano na danych jakościowych i ilościowych. Dążąc do odpowiedzi na założone pytania i przetestowania hipotez przeprowadzono analizę statystyczną za pomocą programu Matlab w wersji R2021b. Korzystając z dostępnych narzędzi przeprowadzono analizę ANOVA wraz z testem Kruskala-Wallis ‘a, testem Shapiro-Wilka oraz analizą Post-Hoc, wykonano estymację parametrów na podstawie przedziałów ufności dla średniej o dużej próbie przy pomocy statystyki testującej i wybranego modelu numer III, zbadano korelację przy pomocy tablicy krzyżowej oraz skorzystano z testu Manna-Whitneya przy testach dla średnich. Przyjęto również za poziom istotności alfa = 0.01, przyjmujemy więc 1% ryzyko na niewystąpienie wykazanych efektów w rzeczywistości, jednak na potrzeby projektu i omawianych danych takie ryzyko jest rozsądne i akceptowalne.

2. Analiza – czy wiek różnicuje średni czas na boisku? [ANOVA]

W celu zbadania wyżej postawionego pytania skorzystano z analizy ANOVA przy uprzednim zweryfikowaniu i dobraniu danych oraz sprawdzeniu założeń wykorzystywanej metody statystycznej. Zestawienie odpowiednio podzielono na grupy wiekowe od 18 do 44 lat w połączeniu ze statystyką średniego czasu spędzonego na boisku. W założeniach odnoszących się do danych sprawdzono:

- Mierzalność analizowanych zmiennych,
- Niezależność zmiennych losowych w rozważanych populacjach,
- Równoliczność analizowanych grup,
- Normalność rozkładu zmiennych w każdej populacji,
- Jednorodność wariancji,

Mierzalność zmiennych spełniona, ponieważ dane wiekowe są wyrażone jako wartość liczbowa.

Niezbędne jest wspomnieć o niezależności zmiennych, gdzie zostało poczynione założenie, iż zawodnicy, którzy otrzymali nagrodę kilkukrotnie nie są traktowani jako wartości zależne, gdyż musimy brać pod uwagę warunkowość względem zmienności ich formy, warunków oraz analizowanego wieku.

Niespełniony został warunek równoliczności analizowanych grup wiekowych, dlatego zgodnie z przyjętymi modelami wykonano test Kruskala-Wallis ‘a, jednak przed przystąpieniem do wykonania sprawdzono pozostałe dwa warunki, które nie zostały spełnione.

Odpowiednio przy weryfikacji normalności rozkładu zmiennych w każdej podgrupie sformułowano poniższe hipotezy:

H0: Dane (w każdej podgrupie) pochodzą z rozkładu normalnego.

H1: Dane (przynajmniej w jednej z podgrupie) nie pochodzą z rozkładu normalnego.

Hipotezy zweryfikowano testem Shapiro-Wilka, jednak poczyniono zawężenie badanych grup, przez wzgląd na jedyne pojedyncze obserwacje w grupach wiekowych 43-44 lata. Odrzucono je, na podstawie założenia, iż przypadki te są nieliczne w całym zbiorze danych, więc nie zmieniają i nie wpływają na analizę. Kontynuowano więc badanie z 25 grupami badanymi, gdzie warunek normalności został spełniony dla grup 18, 40-42 lat. Nie przedstawiona jest analiza p-value ze względu na liczność grup, a decyzyjność zostawiono funkcji napisanej w Matlabie. Tak więc warunek jest niespełniony, są podstawy do odrzucenia hipotezy zerowej i przyjęcia hipotezy alternatywnej.

Group Summary Table

Group	Count	Mean	Std Dev
1	12	9.8333	4.4945
2	143	17.8224	9.5415
3	363	18.8198	9.8447
4	591	19.2267	10.1985
5	1070	18.0299	10.3049
6	1421	17.7602	10.2077
7	1425	19.0632	10.3336
8	1278	20.6634	10.3678
9	1179	21.6718	10.3181
10	1098	22.107	10.1453
11	980	22.8468	9.9439
12	878	22.8754	9.7468
13	815	22.2196	9.7044
14	700	21.6694	9.3014
15	590	20.9007	9.382
16	466	20.5597	9.4649
17	355	20.567	9.2587
18	259	19.9344	9.2604
19	178	19.2713	8.4707
20	117	17.8598	7.9369
21	77	16.7987	7.9947
22	43	15.493	8.7579
23	18	15.3222	8.6836
24	3	14.0667	5.4501
25	3	14.7	2.8
Pooled	14062	20.3924	9.9561

Bartlett's statistic	76.85
Degrees of freedom	24

Rysunek 1. Informacje i parametry statystyczne związane z istotnymi grupami

W przypadku ostatniego warunku, niespełnienie już równoliczności grup, znosi nam wymóg sprawdzenia jednorodności wariancji, dlatego analiza i przedstawienie wyników zostało pominięte.

Kruskal-Wallis ANOVA Table

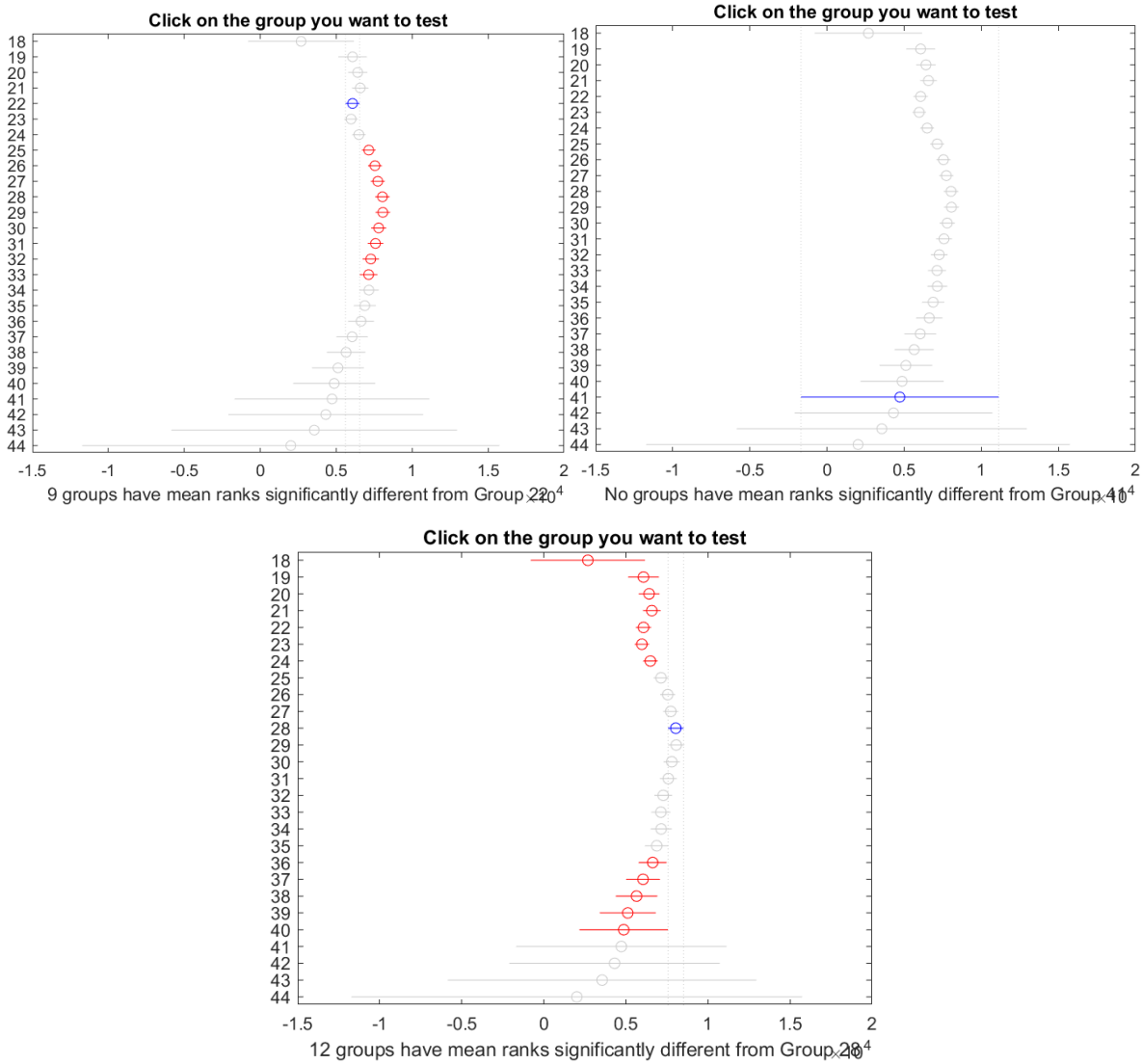
Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	7.81451e+09	26	3.00558e+08	472.18	1.9316le-83
Error	2.25388e+11	14065	1.60247e+07		
Total	2.33202e+11	14091			

Rysunek 2. Wyniki testu Kruskala-Wallis ‘a

Ostatecznie wykonano test Kruskala-Wallis i przyjęto odpowiednie hipotezy:

- Ho: Dane w każdej kolumnie macierzy pochodzą z tego samego rozkładu,
- H1: Nie wszystkie próbki pochodzą z tego samego rozkładu.

Jak widać na powyższym zestawieniu wyników testu Kruskala-Wallis ‘a interesująca wartość p-value przyjęła wielkość równą 1,93e-83, co w porównaniu z przyjętym poziomem istotności 0,01, sprawia, że jest mniejsza od alfy. W takim wypadku odrzucono hipotezę zerową i przyjęto hipotezę alternatywną tj. nie wszystkie próbki pochodzą z tego samego rozkładu. Implikacją takiego wyniku analizy jest wniosek – w odniesieniu do tematu rozważań - iż wiek różnicuje średni czas spędzony na boisku przez zawodników MVP. W wyniku analizy otrzymano, więc wniosek, który można odnieść do dyspozycyjności oraz formy kolejnych zawodników MVP i jest ona zależna od wieku.



Rysunek 3. Wykresy analizy Post-Hoc

W analizie Post-Hoc można zauważyć, iż większość grup wiekowych o dużej liczebności w zakresie od 19 do 30 lat znacznie różnią się od kilku pozostałych grup. Tendencję tą wykazuje większość grup, natomiast te o małej liczebności można traktować jako obserwacje odstające, gdyż rozrzut wartości powodowany jest jedynie

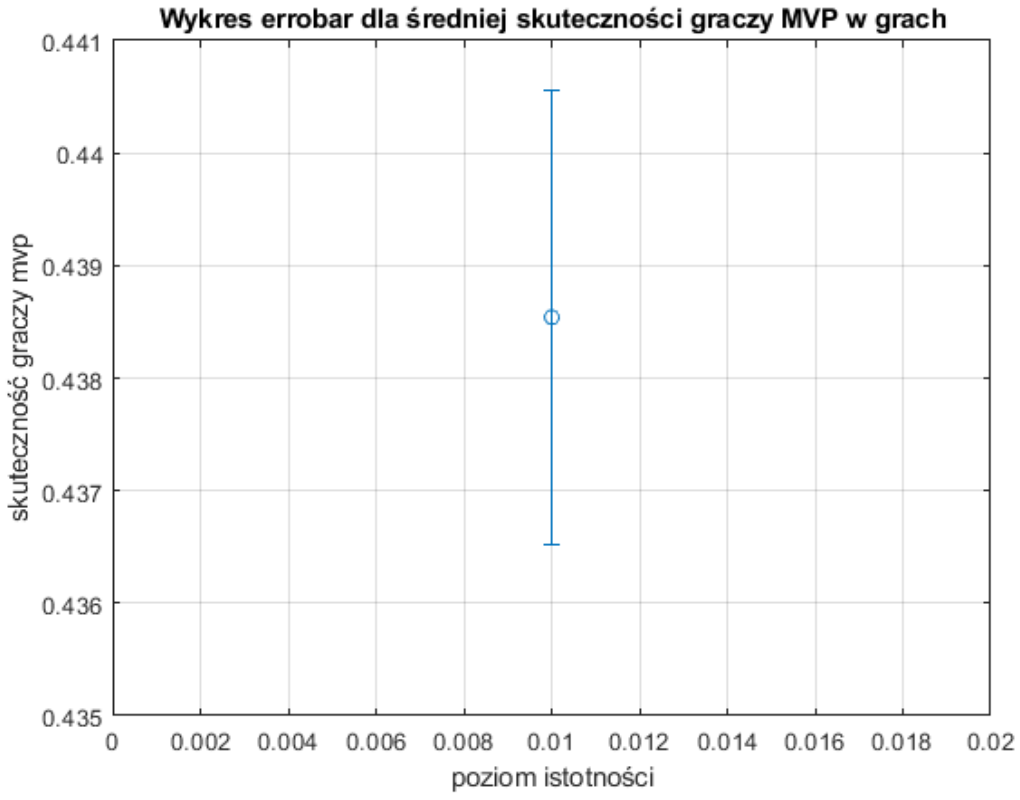
kilkoma obserwacjami w porównaniu do dużych obserwacji. Jednak widząc trend w grupach licznych można wnioskować, iż rozszerzenie grup lub zawężenie zakresu analizy nie zmieniłoby efektów i wyników badania. Można również wysunąć hipotezy na temat określonych grup wiekowych i różnicowania wiekiem m.in. zawodnicy młodsi spędzają mniej czasu na boisku – powodem może być mniejsze doświadczenie i mniej pobłażliwe podejście trenerskie, również i najstarsi spędzają mniej czasu na boisku – tu jednak powodu można doszukiwać się w spadku formy czy wytrzymałości graczy i podejściu "oszczędnym" trenerów do zawodników.

3. Estymacja - skuteczności graczy MVP.

Aby estymować przedziałowo i punktowo skuteczność graczy MVP najpierw zostały usunięte osoby, które nie podały swoich danych na temat skuteczności, a mogły przez to zakłamać wyniki testu. Testy zostały przeprowadzone na poziomie istotności 0.01 dla przedziałów ufności dla średniej z nieznanym odchyleniem standardowym o nieznanym rozkładzie i dla dużej próby, więc wybrany został model trzeci.

Przedział ufności dla średniej z nieznanym odchyleniem standardowym, cecha ma rozkład nieznany, duża próba.			
Estymator	**Statystyka**	**Przedział ufności**	**Kiedy stosować**
Średnia z próby $\bar{x}$, estymator ma rozkład N(m, $\hat{s}/\sqrt{n}$)	$U = \frac{\bar{x} - m}{\sigma}\sqrt{n}$ *statystyka testująca U~N(0,1)*	$P\left\{\bar{x} - u_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}} < m < \bar{x} + u_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}\right\} = 1 - \alpha$ $B(\bar{x}) = \frac{u_{1-\alpha/2}\hat{s}}{\bar{x}\sqrt{n}} * 100\%$	Populacja generalna ma rozkład nieznany **próba duża n>120**.

Za kwantyl przyjęto kwantyl rozkładu normalnego, standaryzowanego. Średnio skuteczność graczy MVP wynosi 43.85%. Dla poziomu ufności 1-alfa=0.99 przedział ufności dla średniej skuteczności graczy MVP w grach wyniósł (43.65, 44.06) przy względnej precyzji szacowania 0.4592%, więc dla dużej precyzji szacowania. Poniżej wykres błędu dla obliczonych wartości.



Rysunek 4. Wykres errorbar dla średniej skuteczności graczy MVP w grach

Obliczona została również estymacja punktowa, czyli błąd standardowy średniej, który wyniósł

7.8181e-04 oraz błąd względny równy 0.1783, zatem obliczona estymacja jest wysoce precyzyjna.

4. Korelacja - zależność pozycji oraz skuteczności, czyli procentowo trafień do kosza w rzutach wolnych.

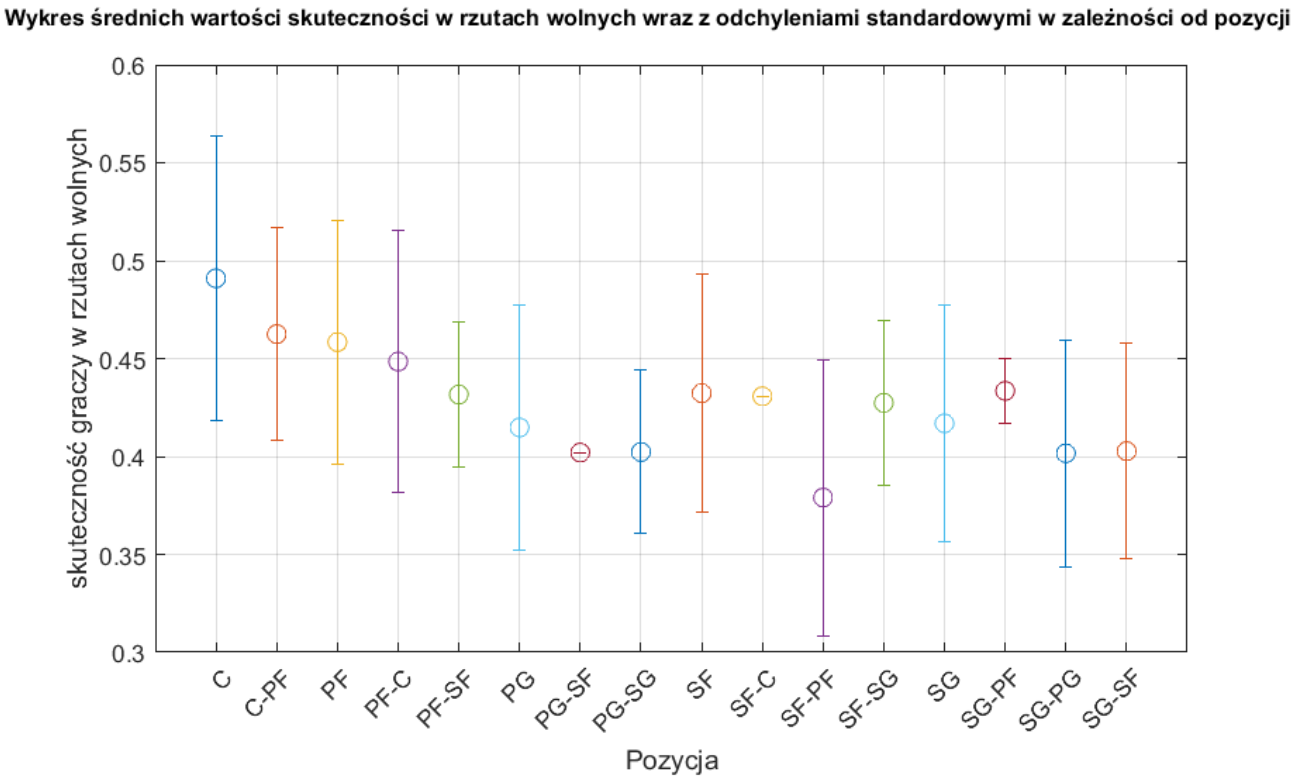
Aby zbadać korelację pozycji i skuteczności, czyli zmienną jakościową i ilościową wymogiem jest obliczenie zależności stochastycznie. Wykorzystany został test chi kwadrat na niezależność zmiennych, a przyjęte zostały hipotezy:

- H0: Cechy statystyczne są niezależne stochastycznie (nie ma zależności pomiędzy pozycją, a skutecznością gracza)

- H1: Cechy statystyczne są zależne stochastycznie (występuje zależność między pozycją, a skutecznością gracza)

```
chi2 = 8.4114e+03
p = 1.5974e-40
```

Jak widać z przeprowadzonego testu wartość p-value wynosi 1.5974e-40, więc jest mniejsza od alfy (0,01). W takim wypadku należy odrzucić hipotezę zerową, czyli cechy są zależne stochastycznie, a w tym przypadku to oznacza, że istotnie jest zależność między pozycją, a skutecznością. Na wykresie została przedstawiona średnia skuteczność w zależności od pozycji.

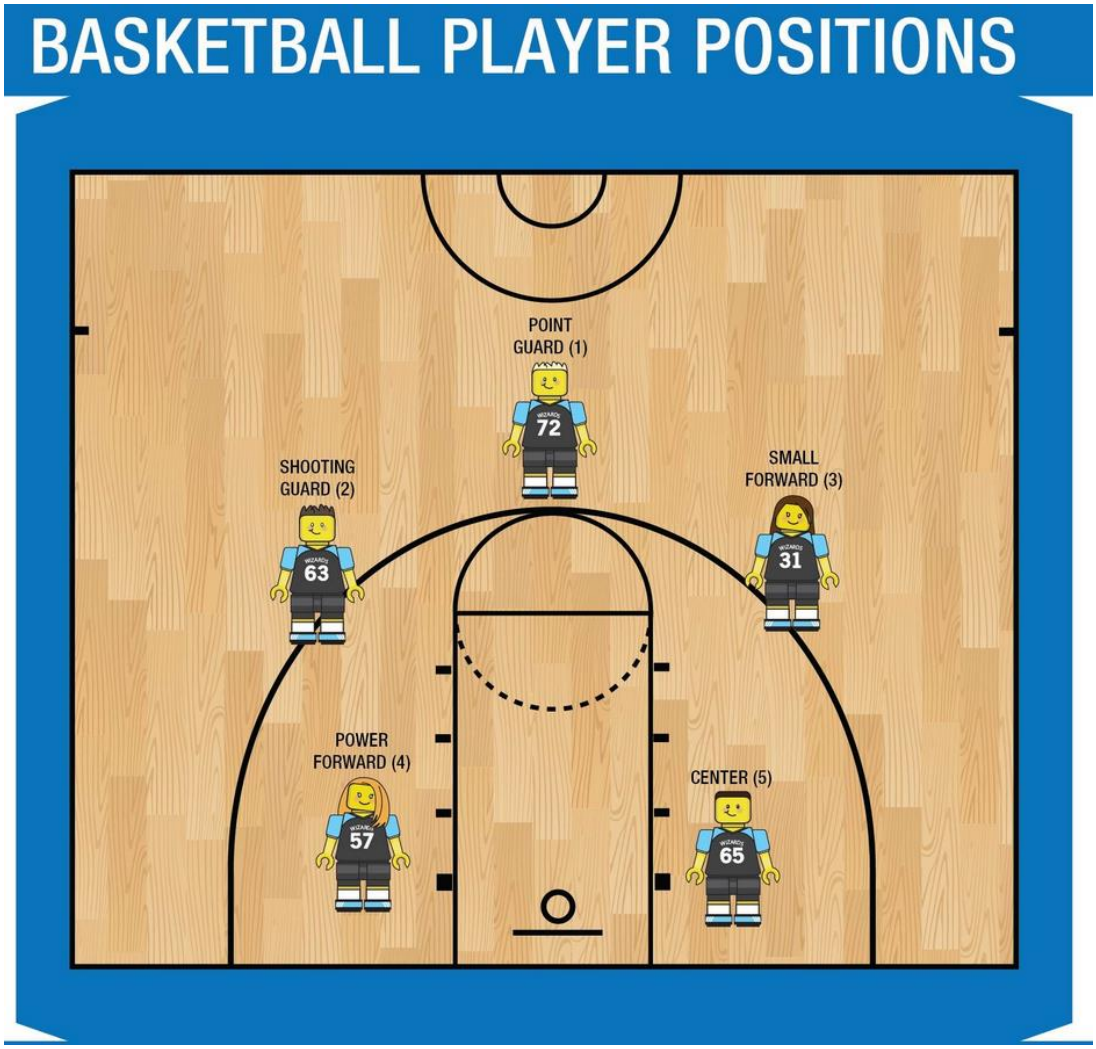


Rysunek 5. Wykresy średnich wartości skuteczności wraz z odchyleniami standardowymi

Jak można zauważyć, średnia skuteczność w większości przypadków oscyluje w granicach 40-50%, lecz

występują spore odchylenia dla niektórych wartości, a nawet zerowe odchylenia. Jest to spowodowane różnymi ilościami danych graczy z różnych pozycji.

Grafika pokazująca rozmieszczenie pozycji graczy na boisku.



Źródło: ready-set-sports.com

5. Porównanie średnich meczy na sezon klubu Los Angeles Lakers względem pozostałych klubów.

Porównujemy ilość meczy, by odnieść wykorzystanie zawodników w klubie i całości populacji (bez danych analizowanego klubu), czyli np. czy w danym klubie można przypuszczać taktykę maksymalnego wykorzystania najlepszych graczy czy oszczędzania ich na najważniejsze momenty lub chwilowe zwyżki formy dające w efekcie nagrodę.

W porównaniu danych meczowych należy rozpocząć od sprawdzenia czy dane pochodzą z rozkładu normalnego z odpowiednimi hipotezami tj.:

H0: Dane (w każdej grupie) pochodzą z rozkładu normalnego.

H1: Dane (przynajmniej w jednej z grup) nie pochodzą z rozkładu normalnego.

Hipotezy zweryfikowano za pomocą testu Shapiro-Wilka. W przypadku danych klubu z Los Angeles, p-value wyniosło 1,11e-16, co sprawia, iż jest to wartość niższa od przyjętego poziomu istotności 0.01. Natomiast w przypadku pozostałych klubów p-value wynosi 0, co oznacza, że wartość również jest niższa od alfy na poziomie 0.01.

Tak, więc w obu przypadkach pojawiają się podstawy do odrzucenia hipotezy zerowej i przyjęcia hipotezy alternatywnej, iż dane nie pochodzą z rozkładu normalnego.

Efektem takiego wyniku poprzedzającego testu jest sprawdzenie czy zmienne są powiązane. Tutaj zgodnie z przyjętą wcześniej konwencją, założono, iż nie są ze sobą powiązane, gdyż baza danych zawiera wyniki klubowe z różnych lat, z różnym składem zawodników, więc dane są poddane czynnikom zmiennej dyspozycyjności, formy zawodników czy warunków klubowych.

Ostatecznie do zweryfikowania naszego tematu skorzystano z testu Manna-Whitneya. W przygotowaniu do przeprowadzenia testu sformułowano odpowiednie hipotezy:

H0: mediana meczy rozegranych w sezonie przez graczy MVP jest taka sama jak w klubie LAL

H1: mediana meczy rozegranych przez LAL jest inna niż rozegranych przez resztę zawodników z klubów

Po przeprowadzeniu analizy otrzymano wartość p-value rzędu 0.26. Jest ona większa od przyjętej wartości alfa = 0.01. Nie ma więc postaw do odrzucenia hipotezy zerowej, iż mediana rozegranych meczy w sezonie przez graczy MVP jest taka sama jak w klubie LAL. W odniesieniu do naszego tematu badawczego wynik analizy uniemożliwia potwierdzenie przypuszczalnych i przedstawionych we wstępie pomysłów, jednak może sugerować podobieństwo w wykorzystaniu zawodników, zachowaniu meczowym i taktyce klubów.

Bibliografia:

- Użyta baza danych:
  https://www.kaggle.com/datasets/vivovinco/19912021-nba-stats?select=player_mvp_stats.csv