Zastosowanie jednostronnego testu ANOVA w celu zbadania zależności odległości sklepów od piekarni w kontekście nieprawidłowości w poziomie zwrotów

Podstawowym narzędziem badania zjawisk jest metoda porównawcza. Polega ona na porównywaniu wielu zbiorów pod kątem pewnej hipotezy. Na przykład ktoś podejrzewa, że w środy występuje największy poziom sprzedaży pieczywa. Oczywiście może podsumować całą roczną sprzedaż chleba w środy i porównać ją z całoroczną sprzedażą w czwartki, piątki itd.

Takie porównanie pokarze ogólną wielkość sprzedaży. To zestawienie nie odpowie jednak na pytanie czy za tak wielką sprzedaż w środę nie odpowiada kilka śród w roku. A może za niższą sprzedaż w czwartki odpowiada kilkanaście przypadków (m.in. tłusty czwartek, wielki czwartek i czwartek weekendu majowego). Pochopne wnioski mogą narazić nas na nieskuteczność lub niepotrzebne wydatki promocyjne. Żyjemy przecież w erze informacji. Pojedyncze podsumowania dają jedynie ogólną i niepraktyczną wiedzę. Aby dowiedzieć się czy środy są statystycznie istotnie lepsze w poziomie sprzedaży powinniśmy porównać zbiór wartości, a nie pojedyncze liczby wynikające z podsumowań.

W tabeli 1 przedstawione jest zestawienie sprzedaży piekarni z 18 tygodni. W pierwszych dwóch kolumnach mamy ilość sprzedaży za poniedziałek i wtorek. Kolejna kolumna, zaznaczona na czerwono, to sprzedaż w środę. Wiersze oznaczają tygodnie sprzedaży. Na pierwszy rzut oka ilość sprzedaży w środę nie różni się od pozostałych dni tygodnia. Ale czy na pewno ta różnica nie jest statystycznie istotna skoro w podsumowaniu rocznym okazało się, że suma sprzedaży w środę jest większa niż w pozostałych dniach tygodnia?

Przykład piekarni Moczary

Piekarnia Moczary sprzedaję swoje pieczywo do 86 okolicznych sklepów wiejskich. Zaopatrywane sklepy znajdują się w promieniu 50 km od piekarni. Piekarnia analizuje niezgodności w ilości zwrotów, rozumiane jako różnica pomiędzy wartością teoretyczną jaka wnika z wielkości dostawy i ilością sprzedaży a faktyczną ilością zwróconego pieczywa. Tygodniowo wartość niezgodności dochodzi nawet do 9%.

Właściciel piekarni zauważył, że w dalej położonych sklepach ilość niezgodności w ilości zwrotów jest wyższa. Właściciel postawił więc hipotezę, że odległość sklepu od piekarni jest statycznie nieistotnym czynnikiem wpływającym na poziom nieprawidłowości w zwrotach. Jest to tak zwana hipoteza zerowa h_0. W tej sytuacji hipotezą alternatywną h_1 jest przeciwne stanowisko mówiące, że odległość ma istotnego znaczenia w poziomie nieprawidłowości.

H_0: odległość nie ma znaczenia w poziomie nieprawidłowości

H_1: odległość ma znaczenie w poziomie nieprawidłowości

Tabela 1. Ilość sprzedaży w poszczególnych dniach tygodnia.

```
[165, 168, 174, 121, 124, 124]
[160, 130, 140, 142, 171, 157]
[144, 127, 145, 145, 133, 122]
[159, 160, 144, 167, 145, 138]
[158, 146, 134, 129, 130, 141]
[172, 137, 172, 126, 136, 168]
[121, 139, 156, 145, 170, 172]
[150, 140, 140, 135, 160, 153]
[122, 130, 153, 152, 131, 173]
[144, 156, 132, 171, 149, 155]
[121, 121, 157, 174, 126, 159]
[161, 173, 138, 136, 125, 162]
[163, 152, 122, 122, 160, 162]
[156, 159, 132, 167, 139, 148]
[121, 135, 156, 174, 133, 163]
[141, 132, 125, 127, 125, 168]
[173, 124, 133, 157, 173, 161]
[135, 136, 141, 156, 144, 136]
```

Tabela 2. Zestawienie sprzedaży pieczywa piekarni Moczary

Nazwa_wsi	Zwroty	Zwroty/dzień	Poziom_niezgodności_zwrotów	Odległość_km
Anielin	37	5,29	0,0151	21,8
Arkowo	51	7,29	0,011	6,6
Alerpin	13	1,86	0,0591	16,2
Baskowice	46	6,57	0,0152	35,1
Halęcin	69	9,86	0,0122	32
...	...	...	...	...
Złodzieje Kolonia	25	3,57	0,0196	17,1
Szkarady	13	1,86	0,0484	19,9
Szkarady Nadzieja	62	8,86	0,0147	17,4
Tupadły	47	6,71	0,0134	8,2
Kapitany Stare	49	7	0,01	53,9

Właściciel podzielił 86 sklepów do przedziałów odległości stosując formułę kwantylową. Poniższa formuła napisana jest w języku programowania Python.

```
df['Grupa_odległości']=pd.qcut(df['Odległość_km'], 4,labels=["do 12 km","12-20 km","20-28 km","28-54 km"])
```

Tabela 3. Poziom niezgodności zwrotów i grupa odległości od piekarni.

Nazwa_wsi	Poziom_niezgodności_zwrotów	Grupa_odległości
Anielin	0.0151	20-28 km
Arkowo	0.0110	do 12 km
Alerpin	0.0591	12-20 km
Baskowice	0.0152	28-54 km
Halęcin	0.0122	28-54 km
...	...	...
Złodzieje Kolonia	0.0196	12-20 km
Szkarady	0.0484	12-20 km
Szkarady Nadzieja	0.0147	12-20 km
Tupadły	0.0134	do 12 km
Kapitany Stare	0.0100	28-54 km

Z uzyskanych wyników utworzono tabelę przestawną.

```
pd.pivot_table(df, index= ['Grupa_odległości'], values= "Odległość_km", aggfunc= ['min','max', 'count'])
```

Tabela 4. Grupy sklepów według odległości od piekarni Moczary.

Grupa_odległości	min Odległość_km	max Odległość_km	count Odległość_km
do 12 km	0.0	12.0	22
12-20 km	12.6	19.9	21
20-28 km	20.2	28.4	21
28-54 km	29.2	53.9	22

Sklepy podzielono na cztery grup; w każdej jest 21-22 sklepy. Wybór ilości grup został przeprowadzony na podstawie subiektywnej decyzji badacza.

Test ANOVA

Analiza wariancji, ANOVA (od ang. *analysis of variance*) jest metodą statystyczną służąca do badania zależności od jednego lub wielu działających równocześnie czynników. Metoda ta pozwala wyjaśnić, czy wskazane czynniki mogą być powodem różnic między obserwowanymi grupami.

W naszym przypadku mamy cztery grupy sklepów według odległości. Stawiamy hipotezę h_0, że odległość nie ma statystycznie istotnego wpływu na poziom nieprawidłowości w procesie zwrotów.

Stosujemy funkcje języka Python:

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

model_H = ols('Poziom_niezgodności_zwrotów ~ C(Grupa_odległości)', data=df).fit()

anova_table = sm.stats.anova_lm(model_H, typ=2)
print(anova_table)
```

Otrzymujemy wartość testu ANOVA:

	sum_sq	df	F	PR(>F)
C(Grupa_odległości)	0.005155	3.0	6.746977	0.0004
Residual	0.020883	82.0	NaN	NaN

Zaznaczona na czerwono wartość wskazuje jaki jest poziom prawdopodobieństwa, że odległość piekarni od sklepów nie ma znaczenia. Wartość prawdopodobieństwa spełnienia hipotezy zerowej, $p < 0,05$ jest bardzo mała co oznacza, że istnieje statystycznie istotny wpływ odległości na poziom nieprawidłowości w zwrotach. Hipoteza zerowa h_0 postawiona na początku badania została więc odrzucona na rzecz hipotezy alternatywnej h_1.

Testu HSD Tukeya

ANOVA pokazała, że istnieje statystycznie istotna różnice w poziomie nieprawidłowości w grupach. Niestety ANOVA nie wskazuje, które grupy odległości istotnie różnią się od siebie. Aby poznać statystyczną istotność różnic w grupach należy przeprowadzić analizę wielokrotnego porównania par (porównanie post-hoc) używając testu HSD Tukeya.

```
from statsmodels.stats.multicomp import pairwise_tukeyhsd

m_comp = pairwise_tukeyhsd(endog=df['Poziom_niezgodności_zwrotów'], groups=df['Grupa_odległości'], alpha=0.05)
print(m_comp)
```

Tabela 5. Tabela istotności testu HSD Tukeya.

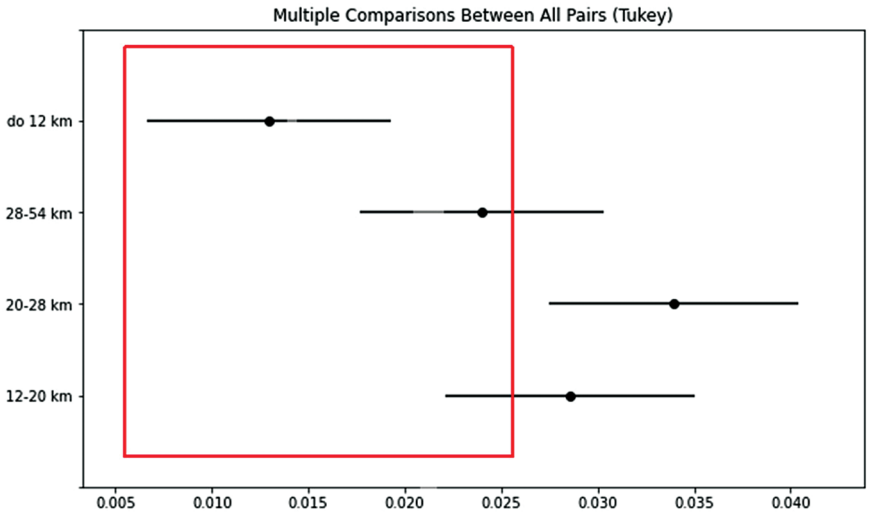
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
12-20 km	20-28 km	0.0054	0.6703	-0.0075	0.0183	False
12-20 km	28-54 km	-0.0046	0.7596	-0.0173	0.0082	False
12-20 km	do 12 km	-0.0156	0.0103	-0.0284	-0.0028	True
20-28 km	28-54 km	-0.01	0.1788	-0.0227	0.0028	False
20-28 km	do 12 km	-0.021	0.001	-0.0338	-0.0082	True
28-54 km	do 12 km	-0.011	0.1087	-0.0236	0.0016	False

Zaznaczone dwie pary odległości wskazują, na istotne statystycznie różnice. Test wskazał na dwie pary:
12-20 km – do 12 km
20-28 km – do 12 km
Tym samym test wskazał, że średnie w pozostałych parach odległości nie różnią się statystycznie istotnie i nie warto się nimi zajmować. Właściciel piekarni otrzymał informację, że trzeba przyjrzeć się sklepom znajdującym się w odległości powyżej 12 km i poniżej 28 km. Lepiej widać te wnioski na poniższym wykresie.

```
m_comp.plot_simultaneous()
plt.vlines(x=20000,ymin=-0.5,ymax=3.5, color="red", alpha=0.8, linestyle='--')
```

Wykres 1. Test HSD Tukeya dla czterech grup odległości.



Na wykresie 1 widać, że przedział nieprawidłowości w przedziałach odległości o *12 km* oraz *28-54 km* jest statystycznie zbliżony. Obszar podobieństw został zaznaczony czerwoną ramką. Oznacza to, że w parze przedziałów *12 km – 28-54 km* została zachowana hipoteza zerowa h_0. Na osi X wykresu zaznaczony jest zakres nieprawidłowości w liczbie zwrotów. Należy zwrócić uwagę na przedziały odległości *20-28* km oraz *12-20 km*, które istotnie różnią się od przedziału *do 12 km*.

Weryfikacja zgodności testu jednorodności wariancji ANOVA

Aby być pewnym wyników testu jednostronnego ANOVA należy zweryfikować podstawowe warunki stosowania tego testu. Są one następujące:

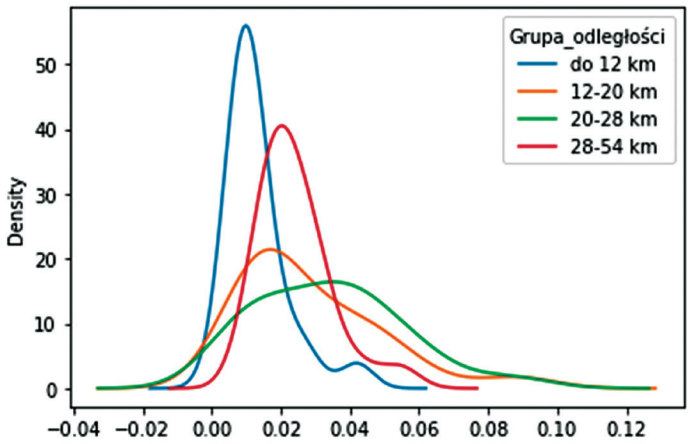
1. wartości rezydualne mają rozkład normalny (test Shapiro Wilksa),
2. wariancje w grupach są jednorodne (test Levene lub Bartlett),
3. obserwacje są prowadzone niezależnie od siebie.

Test Levene'a – sprawdzenie jednorodności wariancji

W tym teście sprawdzamy czy grupy odległości mają jednorodne wariancje. Rozkłady normalne poszczególnych grup powinny być do siebie wizualnie zbliżone. Wtedy prawdopodobnie test Levene'a wykaże spełnienie warunku jednorodności wariancji. Stawiamy więc hipotezę zerową h_0, mówiącą o tym, że wariancje grup odległości od piekarni mają podobną wariancję.

```
PKS = pd.pivot_table(df, index = 'Nazwa_wsi', columns = 'Grupa_odległości', values='Poziom_niezgodności_zwrotów')
P01=PKS['do 12 km'].dropna(how='any')
P02=PKS['12-20 km'].dropna(how='any')
P03=PKS['20-28 km'].dropna(how='any')
P04=PKS['28-54 km'].dropna(how='any')
PKS.plot.kde()
```

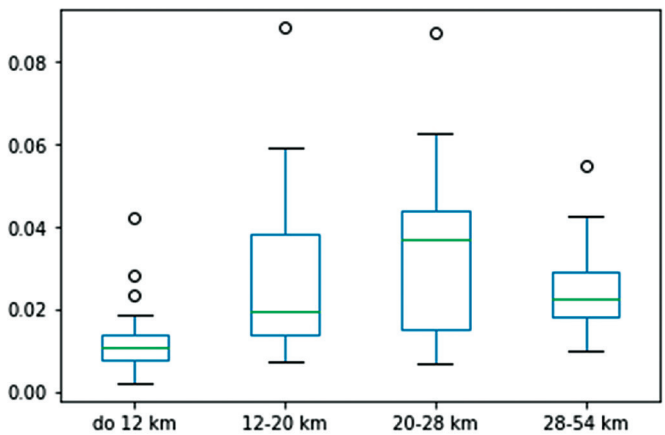
Wykres 2. Rozkłady gęstości prawdopodobieństwa czterech grup odległości.



Jak widać grupa *do 12 km* znacznie się różni od pozostałych. Wykres 2 mówi, że proces zbioru zwrotów jest istotnie najlepszy w sklepach znajdujących się *do 12 km* od piekarni. Niewątpliwie najgorszy przebieg procesu ma miejsce w sklepach oddalonych od piekarni w odległości między 20 a 28 km. Potwierdza to również wykres pudełkowy poniżej.

```
PKS.boxplot(column=['do 12 km', '12-20 km', '20-28 km', '28-54 km'], grid=False)
```

Wykres 3. Wykres grup odległości typu boxplot.



Czas więc przeprowadzić test Test Levene'a.

```
import scipy.stats as stats
w,p = stats.levene(P01,P02,P03,P04)
print("Value:   ",w)
print("p-value: ",p)
```

```
Value:    4.909693810347966
p-value:  0.0034594148761172058
```

Jak widać prawdopodobieństwo p-value jest mniejsze niż 0,05 co oznacza, że należy odrzucić hipotezę zerową h_0 mówiącą o tym, że wariancje grup są jednorodne. Skoro wariancje nie są jednorodne, można postawić w wątpliwość wynik podobieństwa par testu ANOVA.

Test Shapiro-Wilk – sprawdzenia normalności rozkładu reszt

Test ten odpowie na pytanie czy różnice pomiędzy wartościami rezydualnymi i empirycznymi mają rozkład normalny. W tym konkretnym przypadku chodzi o to, aby potwierdzić nieprzydatność testu ANOVA. Wątpliwość taka pojawiła się po teście Levene'a.

```
w, p = stats.shapiro(model_H.resid)
print("Value:   ",w)
print("p-value: ",np.round(p, decimals=2))
```

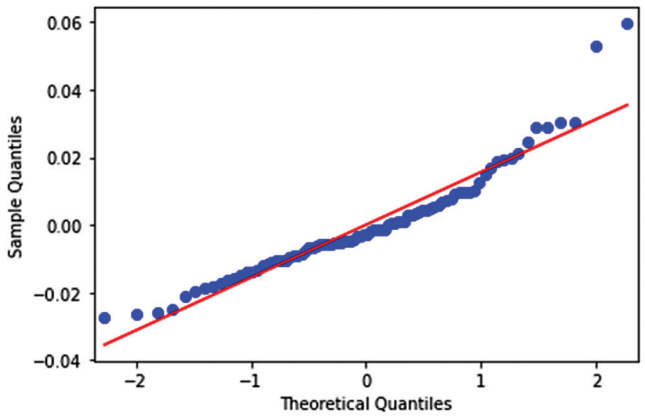
```
Value:    0.9254999756813049
p-value:  0.0
```

Test Shapiro-Wilka pokazał, że prawdopodobieństwo normalności rozkładu reszt nie jest normalne. Niestety test pokazał, że wybrane grupy są niestatystyczne i aby lepiej zbadać zjawisko należy wyznaczyć więcej grup odległości niż 4.

```
from statsmodels.graphics.gofplots import qqplot
from matplotlib import pyplot

qqplot(model_H.resid, line='s')
pyplot.show()
```

Wykres 4. Weryfikacja rozkładu normalnego reszt.



Wykres 4 normalności reszt wskazał, że istnieje kilka sklepów, które statystycznie istotnie odbiegają od przeciętnej. Sklepy te symbolizowane są przez granatowe kropki znajdujące się daleko od czerwonej linii symbolizującej idealny rozkład normalny. Znalezienie tych sklepów powinno istotnie ograniczyć zjawisko nieprawidłowości w prowadzeniu zwrotów.

Wykres pokazał również, że rozkład normalny jest raczej zachowany co oznacza, że nie dochodzi do fałszerstw w prowadzeniu dokumentacji zwrotów. Wykres ten sprawdza się świetnie do weryfikacji czy proces zapisywania zjawiska przez personel sklepów ma prawdziwy charakter, czy może jest prowadzony w sposób kreatywny. Ponieważ testy Shapiro-Wilk oraz Levene'a wyszły nie najgorzej warto przeprowadzić cały proces jeszcze raz z użyciem większej ilości grup odległości.

Wnioski końcowe

Ktoś kto chce zarządzać skutecznie swoją piekarnią musi mieć dokładne informacje. Doświadczony statystyk może szybko przeprowadzić dokładną analizę, zlokalizować miejsca i charakter nieprawidłowości. Oczywiście można też samemu wyliczyć średniotygodniową różnicę inwentaryzacyjną dla każdego sklepu, przesortować wyniki, wybrać najgorsze sklepy i podjąć działania naprawcze. Jednak jeżeli ktoś chce wprowadzić w swojej piekarni modele prognozowania *machine learning* lub chce wprowadzić nowoczesne algorytmy sztucznej inteligencji musi przeprowadzić wyrafinowane proces badawcze. Próbkę takich badań przedstawiłem w tej publikacji.

Wojciech Moszczyński

Wojciech Moszczyński – absolwent Katedry Ekonometrii i Statystyki Uniwersytetu Mikołaja Kopernika w Toruniu, specjalista z zakresu ekonometrii, *data science* oraz rachunkowości zarządczej. Specjalizuje się w optymalizacji procesów produkcyjnych i logistycznych. Prowadzi badania w obszarze rozwoju i zastosowania sztucznej inteligencji. Od lat zaangażowany w popularyzację ekonometrii oraz *data science* w środowiskach biznesowych