Bootcamp Data Science
Zajęcia 3

Przemysław Spurek

Testy Normalności.

Testowanie założenia o normalności rozkładu

We wszystkich wspomnianych powyżej klasycznych testach statystycznych istotnym założeniem jest to, że dane wejściowe w próbie mają rozkład normalny.

Testowanie założenia o normalności rozkładu

We wszystkich wspomnianych powyżej klasycznych testach statystycznych istotnym założeniem jest to, że dane wejściowe w próbie mają rozkład normalny.

W powyższych zadaniach po prostu to zakładaliśmy, ale w praktyce, kiedy dostajemy próbę do analizy, musimy sami sprawdzić, czy możemy uznać ją za pochodzącą z rozkładu normalnego. Do weryfikacji takiej hipotezy służą testy i narzędzia graficzne.

Testowanie założenia o normalności rozkładu

https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z01.ipynb

QQ-Plot

Q w QQ-Plot oznacza kwantyl. Na wykresie kwantyle określonego zbioru danych są zaznaczane na podstawie kwantyli rozkładu referencyjnego, zazwyczaj standardowego rozkładu normalnego.

PP-Plot

Na wykresie przedstawiona jest dystrybuanta CDF (dystrybuanta empiryczna dla zbioru danych) oraz dystrybuanta CDF rozkładu referencyjnego.

Probability Plots

Na wykresie przedstawiona jest posortowana próbka w odniesieniu do kwantyli rozkładu referencyjnego.

Testy Normalności

W testach dotyczących normalności mogą pojawić się różne wyzwania:

- czasami może być dostępnych tylko kilka próbek,
- czasami danych jest bardzo dużo,
- niektóre zawierają bardzo dużo wartości odstających.

Aby sprostać różnym sytuacjom opracowano różne testy normalności. Można je podzielić na dwie kategorie:

- Testy w oparciu o porównanie ("najlepsze dopasowanie") z danym rozkładem, często określane w kategoriach dystrybuanty (CDF). Przykładami są: test Kołmogorowa-Smirnowa, test Lillieforsa, test Andersona Darlinga, kryterium Cramera-von Misesa, testy Shapiro-Wilka i Shapiro-Francia.
- Testy oparte na statystykach opisowych próbki. Przykładami są: test skośności, test kurtozy, test D'Agostino-Pearsona i test Jarque-Bera.

Testy Normalności

https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z02.ipynb

- Na przykład test Lilliefors, który oparty jest na teście Kołmogorowa-Smirnowa, określa odległość między dystrybuantą empiryczną z próbki, a dystrybuantą rozkładu referencyjnego lub pomiędzy empirycznymi dystrybuantami dwóch próbek.

Testy Normalności

https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z02.ipynb

- Na przykład test Lilliefors, który oparty jest na teście Kołmogorowa-Smirnowa, określa odległość między dystrybuantą empiryczną z próbki, a dystrybuantą rozkładu referencyjnego lub pomiędzy empirycznymi dystrybuantami dwóch próbek.
- Oryginalnego testu Kołmogorowa-Smirnowa nie wolno było używać do próbek o liczności mniejszej od ≤ 300.

Testy Normalności

https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z02.ipynb

- Na przykład test Lilliefors, który oparty jest na teście Kołmogorowa-Smirnowa, określa odległość między dystrybuantą empiryczną z próbki, a dystrybuantą rozkładu referencyjnego lub pomiędzy empirycznymi dystrybuantami dwóch próbek.
- Oryginalnego testu Kołmogorowa-Smirnowa nie wolno było używać do próbek o liczności mniejszej od ≤ 300.
- Test Shapiro-Wilka, który zależy od macierzy kowariancji pomiędzy (order statistics) statystykami próbki, może być stosowany dla mniejszych próbek ≤ 50.

Testy Normalności

`https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z02.ipynb`

- Na przykład test Lilliefors, który oparty jest na teście Kołmogorowa-Smirnowa, określa odległość między dystrybuantą empiryczną z próbki, a dystrybuantą rozkładu referencyjnego lub pomiędzy empirycznymi dystrybuantami dwóch próbek.
- Oryginalnego testu Kołmogorowa-Smirnowa nie wolno było używać do próbek o liczności mniejszej od ≤ 300.
- Test Shapiro-Wilka, który zależy od macierzy kowariancji pomiędzy (order statistics) statystykami próbki, może być stosowany dla mniejszych próbek ≤ 50.
- Polecenie Pythona `stats.normaltest(x)` wykonuje test omnibus D'Agostino-Pearsona. Ten test łączy skośność i kurtozę w celu stworzenia jednolitej globalnej statystyki.

Test Kołmogorowa-Smirnowa

- Test K-S dla jednej próbki (K-S one-sample test) bazuje na maksymalnej różnicy pomiędzy empiryczną dystrybuantą (ECDF), a hipotetyczną dystrybuantą (CDF). Jeżeli statystyka D jest znacząca, wówczas hipoteza głosząca, że analizowany rozkład jest normalny powinna zostać odrzucona.

Test Lillieforsa

- W teście K-S pojawia się dość restrykcyjne założenie, że znana jest dokładna wartość średnia oraz odchylenie standardowe analizowanej populacji - nie są one estymowane z danych. W przeciwnym przypadku wartości statystyki D wyznaczone przez Masseya nie są prawdziwe. Jednakże często nie mamy takiej wiedzy i wykorzystywane wartości średnie wyznaczane są na podstawie dostępnych danych, wówczas test normalności przyjmuje dość skomplikowaną warunkowaną hipotezę - Test Lillieforsa.

Test Shapiro-Wilka

- Test ten bazuje na spostrzeżeniu, iż analizując dopasowanie próbnego zbioru danych do rozkładu normalnego (jego wykresu w q-q plot) jest podobne do zadania liniowej regresji - linia diagonalna jest linią idealnego dopasowania, zaś wszystkie odchylenia od niej są podobne do residuów w zadaniu regresji. I właśnie analizując skalę tych odchyleń można określić jakość dopasowania.
- Autorzy testu rekomendują jego wykorzystanie dla małych zbiorów danych (<20) .

Testy omnibus Jarque-Bera

- Test omnibus Jarque-Bera jest testem opartym o kurtozę i skośność.

Testy dla innych postaci rozkładu

https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z03.ipynb

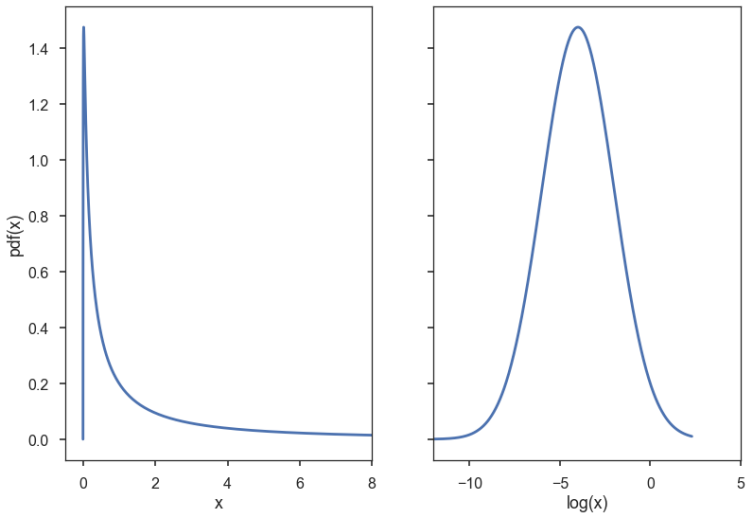
Za pomocą testu Kołmogorowa-Smirnowa możemy zbadać również założenie o pochodzeniu danych z populacji podlegającej dowolnemu innemu rozkładowi ciągłemu.
W tym celu należy podać zamiast 'norm' odpowiednią nazwę rozkładu z modułu `scipy.stats` oraz parametry tegoż rozkładu w odpowiedniej kolejności (należy w tym celu zajrzeć do dokumentacji). Przykład jak to należy zrobić dla rozkładu wykładniczego:

```
D , p = st.kstest(x, 'expon', args=(0, 1/np.mean(x)))
```

A co gdy test nie spełnia założenia o normalności?

Transformacja danych

Jeśli dane znacznie różnią się od rozkładu normalnego, czasami można dokonać transformacji danych tak, by przypominały rozkład normalny. Na przykład dane często zawierają wartości, które zawsze są dodatnie (np. wzrost osób) i które mają długi ogon. W takich przypadkach możemy przetransformować dane stosując logarytm.



Transformacja Boxa-Coxa

Często normalność danych można poprawić przez zastosowanie odpowiedniej transformacji. Ogólną rodzinę transformacji, które często prowadzą do normalizacji danych można zapisać tak: (`http://www.jstor.org/stable/2984418?seq=1#page_scan_tab_contents`):

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{dla } \lambda \neq 0 \\ \ln(y) & \text{dla } \lambda = 0 \end{cases}$$

W module scipy.stats mamy tę transformację zaimplementowaną jako `boxcox()`

Transformacja Boxa-Coxa

https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z04.ipynb

Zadanie

Pobierz dane z pliku.
http://ww2.ii.uj.edu.pl/~spurek/AD_15_16/ex_ad_2.txt oraz:

- Narysuj histogram.
- Sprawdź metodami graficznymi i za pomocą wybranego testu, czy dane pochodzą z rozkładu Normalnego.
- Wykonaj transformatę Boxa-Coxa.
- Sprawdź metodami graficznymi i za pomocą wybranego testu, czy dane po transformacji pochodzą z rozkładu Normalnego.

t-Test Przypomnienie

`https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z05.ipynb`

Załóżmy, że prowadzisz prywatną instytucję edukacyjną. Twoja umowa mówi, że jeśli twoi uczniowie uzyskają 110 punktów w ostatnim egzaminie, gdzie średnia krajowa wynosi 100, to otrzymasz bonus. Jeśli wyniki są znacznie niższe, to tracisz bonus (bo uczniowie nie są wystarczająco dobrzy) i musisz zatrudnić więcej nauczycieli. Jeśli wyniki są znacznie wyższe, tracisz także bonus (ponieważ wydałeś za dużo pieniędzy na nauczycieli) i musisz zredukować liczbę nauczycieli.

Wyniki końcowe egzaminu dla dziesięciu uczniów to:
$[109.4, 76.2, 128.7, 93.7, 85.6, 117.7, 117.2, 87.3, 100.3, 55.1]$.

Pytanie, na które chcemy odpowiedzieć: Czy średnia wartość wyników 97.1 różni się znacząco od wartości 110?

Wilcoxon Signed Rank Sum Test

Jeśli dane nie pochodzą z rozkładu normalnego nie możemy użyć testu t-Studenta.
Zamiast tego musimy użyć testu nieparametrycznego na wartość średnią.
Możemy to zrobić, wykonując test Wilcoxona (test rang).
Test Wilcoxona składa się z trzech kroków:

- Oblicz różnicę między każdą obserwacją, a weryfikowanym parametrem.
- Ignorując znak różnic posortuj je od najmniejszej do największej.
- Oblicz sumę wszystkich ujemnych (lub dodatnich) elementów znajdujących się poniżej lub powyżej wybranego elementu (w zależności od weryfikowanej hipotezy).

Wilcoxon Signed Rank Sum Test

Rozważmy zbiór danych:
$[5260., 5470., 5640., 6180., 6390., 6515., 6805., 7515., 7515., 8230., 8770.]$
oraz hipotezę, że średnia jest równa 7725.

Subject	Daily energy intake (kJ)	Difference from 7725 kJ	Ranks of differences
1	5260	2465	11
2	5470	2255	10
3	5640	2085	9
4	6180	1545	8
5	6390	1335	7
6	6515	1210	6
7	6805	920	4
8	7515	210	1.5
9	7515	210	1.5
10	8230	−505	3
11	8770	−1045	5

Wilcoxon Signed Rank Sum Test

`https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z06.ipynb`

Suma wartości ujemnych daje wartość $3 + 5 = 8$. Wartość odpowiedniego rozkładu można poszukać w tablicach. W praktyce wartość jest obliczana przez komputer:

```
(rank, pVal) = stats.wilcoxon(data-checkValue)
```

Mann–Whitney Test

Jeśli dane nie pochodzą z rozkładu normalnego nie możemy użyć testu t-Studenta.

- Jeśli wartości pomiarów z dwóch grup nie pochodzą z rozkładu normalnego, musimy skorzystać z testu nieparametrycznego.
- Najczęstszym testem nieparametrycznym dla porównania dwóch niezależnych grup jest test Mann-Whitney (Wilcoxon).
- Ten test jest czasami określany również jako test rang Wilcoxona.
- `u_statistic, pVal = stats.mannwhitneyu(group1, group2)`

Zadanie

https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z08.ipynb

Zadanie

Na podstawie danych poniżej przedstawiających wagę losowo wybranych noworodków stwierdzić, czy występuję istotna różnica między średnią wagą chłopców i dziewczynek, oraz czy zróżnicowanie wagi wśród chłopców i dziewczynek jest podobne, czy też różne.

Chłopcy:

3.19, 3.29, 3.31, 3.05, 4.15, 4.26, 3.36, 3.25, 3.18, 2.75, 2.78, 3.22, 3.36, 3.12

Dziewczynki:

3.34, 3.36, 3.22, 3.14, 3.65, 3.15, 3.28, 3.18, 3.35, 3.61, 3.05, 3.03, 3.01, 3.06, 2.9, 3.02

Zadanie

https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z09.ipynb

Zadanie

W ankiecie przeprowadzonej wśród kilkudziesięciu osób zamieszkałych w dwóch regionach pytano m.in. ile razy dziennie piją kawę. Uzyskano następujące odpowiedzi:

Region 1

3 ,4 ,1 ,1 ,4 ,2 ,1 ,1 ,0 ,1 ,4 ,1 ,1 ,6 ,1 ,1 ,0 ,1 ,3 ,1 ,1 ,1 ,2 ,0 ,1 ,0 ,3 ,0 ,1 ,1 ,1 ,0 ,3 ,0 ,0 ,0 ,1 ,2 ,2 ,1 ,1 ,1 ,0 ,0 ,1

Region 2

0 ,0 ,1 ,2 ,0 ,0 ,1 ,1 ,5 ,1 ,1 ,2 ,1 ,2 ,1 ,0 ,1 ,1 ,0 ,0 ,0 ,1 ,0 ,2 ,1 ,2 ,2 ,1 ,1 ,2 ,2 ,6 ,1 ,3 ,2 ,4 ,1 ,2 ,1 ,0 ,4 ,0 ,2 ,0 ,3

Czy na podstawie tych wyników można stwierdzić, że osoby z regionu 1 częściej piją kawę?

Zadanie

https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z10.ipynb

Zadanie

Dla sprawdzenia poprawności działania urządzenia automatycznie odważającego produkt w opakowania po 250 g wybrano losowo 15 opakowań i zważono masę netto każdego z opakowań. Uzyskano następujące wyniki:
249.3, 248.5, 250.2, 249.3, 247.9, 250.3, 251.1, 250.2, 249.3, 248.3, 247.9, 248.6, 250.6, 251.6, 249.6
Czy na podstawie uzyskanych wyników możemy stwierdzić, że urządzenie odważa średnio właściwą masę produktu?

ANOVA

Analysis of Variance (ANOVA)

Pomysł analizy wariancji (ANOVA) polega na podzieleniu wariancji na:

- wariancję między grupami (variance between groups),
- wariancję wewnątrz grup (variance within groups),

oraz sprawdzeniu czy te rozkłady odpowiadają hipotezie zerowej: Grupy pochodzą z tego samego rozkładu.
Na podstawie wyników w próbie należy zweryfikować hipotezę:

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

względem hipotezy alternatywnej

$$H_1 : \mu_i \neq \mu_j, \text{ gdzie } i \neq j.$$

Analysis of Variance (ANOVA)

Na przykład, jeśli porównamy grupę bez leczenia z grupami odpowiednio:

- z leczeniem A,
- z leczeniem B,

wykonujemy jednoczynnikową analizę wariancji (ANOVA), czasami zwaną jednokierunkową.
Jeśli wykonamy taki sam test na mężczyznach i kobietach, to mamy dwuczynnikową analizę wariancji (ANOVA). Względem płci oraz typu leczenia.
Zauważ, że w analizie wariancji (ANOVA) ważne jest aby mieć dokładnie taką samą liczbę próbek w każdej z grup!

Analysis of Variance (ANOVA)

- Ponieważ hipoteza zerowa mówi, że nie ma żadnej różnicy między grupami, test opiera się na porównaniu różnic miedzy próbkami (tzn. pomiędzy ich środkami).
- Porównanie przyjmuje ogólną formę testu-F w celu porównania wariancji, ale w przypadku dwóch grup test t-Studenta prowadzi do dokładnie takiego samego wyniku.
- Jednoczynnikowa ANOVA zakłada, że wszystkie próbki pochodzą z rozkładu normalnego o tej samej wariancji. Założenie równej wariancji można sprawdzić przy użyciu testu Levene.

Jednoczynnikowa analiza wariancji

`https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z11.ipynb`

Jako przykład rozważmy dane zawierające poziomy przepływu ($\mu g/l$) w trzech grupach pacjentów z zastawką serca, z różnymi poziomami wentylacji azotem. W analizie wzięło udział 22 osoby.
Zerowa hipoteza dla ANOVA mówi, że wszystkie grupy pochodzą z tej samej populacji.

Jednoczynnikowa analiza wariancji

https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z11.ipynb

Jako przykład rozważmy dane zawierające poziomy przepływu ($\mu g/l$) w trzech grupach pacjentów z zastawką serca, z różnymi poziomami wentylacji azotem. W analizie wzięło udział 22 osoby.
Zerowa hipoteza dla ANOVA mówi, że wszystkie grupy pochodzą z tej samej populacji.
Bardziej szczegółowy wynik ANOVA można otrzymać za pomocą modelowania statystycznego:

	DF	SS	MS	F	p(>F)
C(treatment)	2	15515.76	7757.88	3.71	0.043
Residual	19	39716.09	2090.32	NaN	NaN

Jednoczynnikowa analiza wariancji

- Najpierw oblicza się Sums of squares (SS). Otrzymujemy $SS_{Treatments} = 15515.76$ oraz $SS_{Error} = 39.716.09$.
- Średnie kwadraty (mean squares) (MS), to SS podzielone przez odpowiednie stopnie swobody (DF).
- Otrzymujemy również wartość statystyki F:

$$F = \frac{MS_{Treatments}}{MS_{Error}} = \frac{SS_{Treatments}/(n_{groups} - 1)}{MS_{Error}/(n_{total} - n_{groups})}$$

- Otrzymujemy, również p-value

Porównanie kilku grup

- Zerowa hipoteza w jednoczynnikowej analizie wariancji mówi, że wszystkie próbki są takie same. Więc jeśli jednostronnie odrzucimy hipotezę zerową, to nie mamy żadnej informacji.

Porównanie kilku grup

- Zerowa hipoteza w jednoczynnikowej analizie wariancji mówi, że wszystkie próbki są takie same. Więc jeśli jednostronnie odrzucimy hipotezę zerową, to nie mamy żadnej informacji.
- Często nie interesuje nas czy wszystkie próbki są takie same, ale chcielibyśmy też wiedzieć, dla których par próbek takie podobieństwo nie zachodzi.

Porównanie kilku grup

- Zerowa hipoteza w jednoczynnikowej analizie wariancji mówi, że wszystkie próbki są takie same. Więc jeśli jednostronnie odrzucimy hipotezę zerową, to nie mamy żadnej informacji.
- Często nie interesuje nas czy wszystkie próbki są takie same, ale chcielibyśmy też wiedzieć, dla których par próbek takie podobieństwo nie zachodzi.
- Analiza takich zależności nazywana jest porównaniami **post hoc** lub testami post hoc.

Tukey's Test

Test Tukeya, czasami określany jest jako Honest Significant Difference test (HSD) pokazuje, które średnie różnią się w sposób istotny statystycznie.

- Opiera się ona na statystyce:

$$q_n = \frac{\max\{x_1, \ldots, x_n\} - \min\{x_1, \ldots, x_n\}}{s},$$

  gdzie s jest odchyleniem standardowym z próbki.
- W teście Tukeya próbki $x_1, \ldots, x_n$, to środki grup.
- Test ten może być użyty jako analiza **post hoc** po odrzuceniu hipotezy zerowej w ANOVA, że wszystkie grupy pochodzą z tej samej populacji.

Tukey's Test

https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z13.ipynb

Zadanie

Trzej łucznicy - Patryk, Jacek i Aleksander biorą udział w konkursie strzeleckim. Pierścienie na tarczy mają wartości punktacji od 1 do 10 (10 to najwyższy wynik). Każdy uczestnik strzela 6 razy, zdobywając punkty:
Patryk - 5, 4, 4, 3, 9, 4
Jacek - 4, 8, 7, 5, 1, 5
Aleksander - 9, 9, 8, 10, 4, 10
Na podstawie powyższych wyników chcielibyśmy wiedzieć, kto jest najlepszym łucznikiem. Innymi słowy, nasza hipoteza zerowa oznacza, że średnie wszystkich populacji są jednakowe.

Tukey's Test

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
Aleksander	Jacek	-3.3333	-6.5755	-0.0911	True
Aleksander	Patryk	-3.5	-6.7422	-0.2578	True
Jacek	Patryk	-0.1667	-3.4089	3.0755	False

- Wyniki testu Tukey pokazują średnią różnicę, przedziały ufności i to, czy należy odrzucić hipotezę zerową dla każdej pary grup na danym poziomie istotności.
- W tym przypadku test sugeruje odrzucenie hipotezy o równości średnich dla par:
  - Aleksander Jacek
  - Aleksander Patryk
- To sugeruje, że wyniki Aleksandra stanowczo różnią się od innych grup.
- Wizualizacja 95% przedziałów ufności wzmacnia wyniki w sposób wizualny.

Tukey's Test

https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z14.ipynb

Zadanie

Pewna grupa poddała się trzem testom. Na podstawie wyników (wyniki na stronie) chcielibyśmy zweryfikować hipotezę zerową mówiącą, że średni wynik w tych trzech testach jest taki sam.

Gdy dane nie pochodzą z rozkładu normalnego

`https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z15.ipynb`

Gdy dane nie pochodzą z rozkładu normalnego.

- Kiedy porównujemy dwie grupy ze sobą, używamy testu t-Studenta, gdy dane pochodzą z rozkładu normalnego, a w przeciwnym przypadku testu nieparametrycznego Mann-Whitney.
- W przypadku trzech lub więcej grup testem dla danych z rozkładu normalnego jest analiza wariancji (ANOVA), w odwrotnym przypadku używamy testu Kruskala-Wallisa.
- Gdy hipoteza zerowa jest prawdziwa, to statystyka testu Kruskala-Wallisa ma rozkład chi kwadrat.

2-CZYNNIKOWA ANOVA

- Analiza wariancji (ANOVA) dla klasyfikacji podwójnej bada wpływ dwóch czynników klasyfikujących.
- Zwykle czynniki klasyfikujące oznaczamy wielkimi literami alfabetu łacińskiego: A, B
- Zakładać będziemy, że we wszystkich podgrupach wyznaczonych przez czynniki klasyfikujące znajduje się taka sama liczba obserwacji.
- Za pomocą dwuczynnikowej analizy wariancji testować będziemy zestaw hipotez:

$$H_{A0}: \text{Źródło zmienności A nie różnicuje wyników.}$$

$$H_{B0}: \text{Źródło zmienności B nie różnicuje wyników.}$$

$$H_{AB0}: \text{Źródło zmienności AB nie różnicuje wyników.}$$

2-CZYNNIKOWA ANOVA

https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z16.ipynb

W porównaniu z jednostronnymi ANOVA analiza 2-czynnikowa bada nie tylko pojedyncze czynniki ale może też sprawdzić, czy współdziałanie czynników ma istotny wpływ na rozkład danych.
Weźmy np. pomiary obwodu głowy płodu wykonane przez czterech lekarzy na trzech różnych płodach.

	df	sum_sq	mean_sq	F	PR(>F)
C(fetus)	2	324.00	162.00	2113.10	1.05e-27
C(observer)	3	1.19	0.39	5.21	6.497-03
C(fetus):C(observer)	6	0.56	0.09	1.22	3.29e-01
Residual	24	1.84	0.07	NaN	NaN

Różne płody wykazują znaczne różnice w rozmiarach głowy ($p < 0.001$), również wybór obserwatora ma znaczący wpływ ($p < 0.05$). Jednak żaden indywidualny obserwator nie był znacząco różny przy badaniu każdego indywidualnego płodu ($p > 0.05$).

Zadanie

Wykonaj ANOVA (2-czynnikowa) dla danych:

- https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z17.ipynb
- https://github.com/przem85/bootcamp/blob/master/statistics/D07_Z18.ipynb