Konsultacja statystyczna do publikacji *Brewing antioxidants from spent coffee grounds: the optimization of the extraction process and the comparison of aquatic and ethanolic antioxidant extracts*

Jan Malinowski
(Dated: 30 maja 2025)

I. UWZGLĘDNIANIE NIEPEWNOŚCI POMIAROWYCH

Każdy pomiar w świecie rzeczywistym cechuje się określoną niepewnością pomiarową. Najlepszym sposobem wyznaczenia niepewności pomiarowej wielkości złożonej z większej liczby wyników jest znanie ich indywidualnych niepewności. W sytuacji gdy oszacowanie ich jest problematyczne bądź niemożliwe można skorzystać z metody statystycznej.

Jeśli pomiar jednej wielkości $\boldsymbol{x_i}$ (przy tych samych parametrach) jest wykonywany niezależnie n razy, wówczas za jej najlepsze oszacowanie przyjmujemy wynik $\widehat{x_i} = \overline{x_i} \pm u_{x_i}$, gdzie kolejne zmienne są odpowiednio średnią arytmetyczną n pomiarów oraz jej odchyleniem standardowym, liczonymi za pomocą poniższych wzorów:

$$\overline{x_i} = \sum_{j=1}^{n} \frac{x_{ij}}{n}, \tag{1}$$

$$u_{x_i} = \sqrt{\frac{1}{(n-1)} \sum_{j=1}^{n} (x_{ij} - \overline{x_i})^2}, \tag{2}$$

gdzie x_{ij} oznacza $j$ty pomiar wielkości x_i. W przypadku Twojej pracy pomiarami tej samej wielkości są trzykrotne powtórzenia danej próby. Zakładając, że niepewności pojedynczych pomiarów pochodzą z rozkładu normalnego (należałoby to w jakiś sposób uzasadnić, ale w naukach przyrodniczych bardzo często można to założenie przyjąć z natury pomiarów; dodatkowo pokazanie tego wymagałoby posiadania większej ilości danych) można wówczas dla grupy wyników $\boldsymbol{x_i}$ obliczyć średnią używając niepewności u_{x_i} jako wag w następujący sposób:

$$\overline{x} = \frac{\sum_{i=1} x_i/u_{x_i}^2}{\sum_{i=1} 1/u_{x_i}^2},$$

$$u_{\overline{x}} = \sqrt{\frac{\sum_{i=1}(x_i - \overline{x})^2/u_{x_i}^2}{(N-1)\sum_{i=1}(1/u_{x_i}^2)}},$$

gdzie N jest liczbą różnych pomiarów wielkości x_i. Znalezione w ten sposób wartości można następnie przedstawić w postaci graficznej umieszczając odpowiednie niepewności jako tzw. wąsy na wykresach.

Można oczywiście alternatywnie użyć zwykłej średniej arytmetycznej i policzyć średnie dla grup ze wzorów (1) oraz (2), jednak dla wystarczająco dużych różnych niepewności pomiarowych, wynik będzie mniej dokładny. Dla dokładnie takich samych niepewności pomiarowych u_{x_i} wzory wprost sprowadzają się do pierwszych wzorów przy czym w tym wypadku odchylenie standardowe dla średniej z grupy można policzyć ze wzoru

$$u_{\overline{x}} = \sqrt{\frac{1}{n(n-1)} \sum_{j=1}^{n} (x_i - \overline{x})^2}, \tag{3}$$

jednak kwestia czy należy szacować niepewności dla poszczególnych x_i jest częściowo uznaniowa. W przypadku gdyby te niepewności różniły się warto byłoby to uwzględnić, dla minimalnych bądź braku różnic nie byłoby tak informatywne z racji małej ilości ($n = 3$) powtórzeń dla eksperymentów.

II. TESTOWANIE RÓŻNIC POMIĘDZY GRUPAMI WYNIKÓW

Pierwszym krokiem analizy statystycznej jest postawienie hipotezy testowej. Przyjmuje ona formę zdania logicznego twierdzącego, na które można jednoznacznie odpowiedzieć twierdząco bądź przecząco np. popucja Polski liczy 80 milionów obywateli, co jest zdaniem nieprawdziwym. Zwykle za hipotezę zerową przyjmuje się zdanie, któremu chce się zaprzeczyć. Jest to spowodowane tym, że analiza statystyczna została skonstruowana w taki sposób, że nie pozwala ona na "potwierdzanie"pewnych faktów, a wyłącznie na odrzucanie tych, które nie pasują do wyników na pewnym poziomie istotności. W związku z tym formalnie jeśli chce się uzasadnić prawdziwość pewnego zdania μ_1 należy za hipotezę zerową (a więc tą, którą chcemy w rzeczywistości odrzucić) przyjąć zdanie mu przeciwne μ_0 tak jak to poniżej zilustrowano.

Przykład
μ_0 : Populacja Polski liczy 80 milionów ludzi.
μ_1 : Populacja Polski nie liczy 80 milionów ludzi.

Wówczas tak wybraną hipotezę zerową μ_0 chcemy obalić poprzez użycie wybranego testu statystycznego.

Samo postawienie hipotezy nie jest jednak wystarczające. W celu sprawdzenia jej (nie)prawdziwości potrzebne jest określenie poziomu istotności α. Parametr ten zwykle ustala się na poziomie 0.05 (stąd słynne $p < 0.05$) i informuje w jakiej części testów możemy błędnie odrzucić hipotezę zerową. W tym konkretnym wypadku oznaczałoby to, że przeprowadzając dowolny test mamy 5% szansą na to, że błędnie odrzucimy hipotezę zerową pomimo, że była ona prawdziwa. Jest to

ryzyko, którego nie da się zlikwidować, ale zwykło się przyjmować, że jeśli obliczony poziom istotności p dla statystyki z danego testu będzie niższy od 0.05 to można hipotezę zerową odrzucić (przy czym wciąż jest to wybór arbitralny, ale tak jest skonstruowany aparat analizy statystycznej). Posiadając ustalony poziom istotności α oraz hipotezę zerową, można zabrać się do wyboru testu statystycznego.

A. Test *t* Studenta

Spośród wielu testów statystycznych, test t Studenta jest jednym z najbardziej popularnych. Zakładając, że dane testowane pochodzą z rozkładu normalnego, pozwala on stwierdzić czy różnica pomiędzy średnimi wynikami z dwóch wybranych grup wyników jest istotna statystycznia, a co za tym idzie czy wyniki w tych grupach pochodzą z różnych rozkładów. Nie ma znaczenia w tym wypadku jak liczne są grupy wyników (chociaż im więcej danych tym można liczyć na dokładniejszy wynik) i ich liczności nie muszą być takie same. Istotne jest natomiast sformułowanie właściwej hipotezy bo można zbadać zarówno czy średnie są od siebie różne jak i czy, któraś z nich jest konkretnie mniejsza (postawienie hipotezy w taki sposób modyfikuje nieco zakres obszaru, w którym będziemy odrzucać hipotezę zerową).

W celu przeprowadzenia testu t Studenta należy wybrać dwie grupy obserwacji o licznościach N_1 oraz N_2, co do których istnieje podejrzenie, że pochodzą z różnych rozkładów. Hipotezą zerową jest w tym wypadku stwierdzenie, wyniki w obydwu próbach pochodzą z tej samej populacji. Dla każdej z tych dwóch grup należy obliczyć średnie $\overline{x_1}$, $\overline{x_2}$ oraz wariancje $u^2_{x_1}$, $u^2_{x_2}$ korzystając po prostu ze wzorów (1) i (2). Otrzymane wyniki można następnie wykorzystać do policzenia estymaty wariancji różnicy wartości średnich zgodnie z poniższym wzorem

$$u^2_\Delta = \frac{N_1+N_2}{N_1N_2}\frac{(N_1-1)u^2_{x_1}+(N_2-1)u^2_{x_2}}{N_1+N_2-2}.$$

Tak obliczoną wielkość można finalnie wykorzystać do znalezienia wartości statystyki t, która po porównaniu z rozkładem wskaże poziom istotności wyniku.

$$t = \frac{\overline{x_1}-\overline{x_2}}{u_\Delta}.$$

Aby określić poziom istotności otrzymanego wyniku należy odczytać z tablic rozkładu t o $f = N_1+N_2-2$ stopniach swobody wartości graniczne (dla $2,5\%$ z obydwu stron, bądź 5% z jednej strony dla testu jednostronnego - gdy pytamy o to czy jedna średnia jest mniejsza od drugiej) i porównać je z otrzymaną wartością t. Jeśli znajduje się ona w przedziałach dopuszczalnych przy danym poziomie istotności to hipotezy zerowej nie można odrzucić.

Test t Studenta może być wykonany za pomocą różnych pakietów programowania, bądź przeprowadzony ręcznie obliczając wszystkie wymienione powyżej wielkości. Podczas używania zaimplementowanych metod testu t należy zwrócić uwagę na 3 istotne parametry:

- zwrócenie uwagi czy dana implementacja testu jest jedno czy dwustronna (hipoteza o różnych średnich vs mniejszej/większej),
- wpisanie właściwej liczby stopni swobody (zwykle jeśli wejściem dla pakietu są grupy danych to jest to robione automatycznie),
- określenie poziomu istotności testu.

Uwaga o wielokrotnym testowaniu hipotez

Przyjęcie poziomu istotności α na poziomie 0.05 oznacza, że mamy szansę 5%, że odrzucimy hipotezę, która była prawdziwa, a więc przyjmiemy za prawdziwą fałszywą hipotezę alternatywną. Można na podstawie tego wnioskować, że im więcej testów zostanie wykonanych, tym więcej testów przypadkowo będzie wskazywało na przyjęcie niewłaściwej hipotezy. W związku z tym najlepszym sposobem byłoby przeprowadzenie tylko tych testów odnośnie, których ma się solidne przypuszczenia, że mogą wskazać na istotny wynik, ale zwykle trudno zawczasu stwierdzić, które wyniki takie się okażą. Rozwiązaniem tego problemu jest zastosowanie poprawek poziomu istotności pojedynczych testów.

W przypadku testów wielokrotnych chcemy kontrolować wielkość FWER - prawdopodobieństwo popełnienia przynajmniej jednego błędu I rodzaju w grupie testów (jest to błąd, za który odpowiadał poziom istotności indywidualnego testu). W celu utrzymania tej wielkości na poziomie istotności $p = 5\%$ możemy skorzystać z następujących poprawek:

- poprawki Bonferroniego, w której dla każdej z m testowanych hipotez dobieramy poziom istotności $\frac{\alpha}{m}$ - jest to dosyć mocna poprawka, ponieważ wymaga bardzo silnie istotnych wyników,
- poprawki Bonferroniego-Holma, w której każdy test będzie wykonany na poziomie istotności $\frac{\alpha}{m+1-k}$, gdzie k należy znaleźć korzystając z poniższego algorytmu:

1. obliczyć poziomy istotności p_i dla każdej hipotezy,
2. uporządkować je w kolejności rosnącej tak żeby $p_{(1)} < p_{(2)} < ... < p_{(m)}$,
3. dla określonego poziomu FWER $= \alpha$ (przyjmiemy tutaj raczej 0.05) znaleźć najmniejsze k, dla którego zajdzie warunek
$$p_k > \frac{\alpha}{m+1-k},$$
4. odrzucić hipotezy $\mu_1, \mu_2, ..., \mu_{k-1}$, przyjmując $\mu_k, ..., \mu_m$.

- niektóre prace krytykują utrzymywanie wskaźnika FWER na pewnym określonym poziomie, wskazując, że nie jest tak istotne aby nie popełniać żadnego błędu I rodzaju. Sugerowanym podejściem jest wówczas kontrolowanie wskaźnika FDR (False Discovery Rate), który jest oczekiwanym stosunkiem liczby błędów I rodzaju do wszystkich odrzuconych hipotez zerowych, pomnożony przez prawdopodobieństwo odrzucenia co najmniej jednej hipotezy. Jednym ze sposobów kontrolowania FDR na poziomie δ (zwykle można przyjmować podobnie jak α na poziomie 5%) jest metoda Benjamini i Hochberga, której schemat zaprezentowano poniżej:

5. należy obliczyć poziomy istotności p_i dla każdej hipotezy oraz uporządkować je w kolejności rosnącej tak jak w poprawce Bonferroniego-Holma,

6. znaleźć największą liczbę j, dla której zajdzie nierówność

$$p_j \leq \delta \frac{j}{m},$$

7. przyjąć wszystkie hipotezy $\mu_1, \mu_2, ..., \mu_j$ za istotne.

- użycie metody Storeya do kontroli wskaźnika pFDR (positive False Discovery Rate), który dla dużej liczby testowanych hipotez jest równoważny wartości FDR i określa oczekiwany stosunek liczby błędów I rodzaju do wszystkich określonych hipotez zerowych. W tym wypadku warto użyć jakiegoś pakietu oprogramowania.

Oczywiście można zrezygnować z użycia poprawki, ale dla dużej ilości testów ryzykuje się wtedy przyjęcie niektórych hipotez zupełnie przypadkowo. Generalnie zasada wygląda tak, że trudno powiedzieć zawczasu, której poprawki powinno się użyć. Warto przeliczyć poziomy istotności dla wielu hipotez i zobaczyć ile hipotez należałoby przyjąć/odrzucić stosując konkretną poprawkę, a następnie zdecydować się na użycie tej, która odpowiada naszym przekonaniom. Bezpiecznie możnaby np. próbowac użyć poprawki Bonferroniego-Holma czy procedury Benjamini i Hochberga.

Metody pozwalające sprawdzić normalność rozkładu wyników

W poprzednich sekcjach wspomniane było, że czasami istotnym założeniem przy przeprowadzaniu pewnych obliczeń jest pochodzenie danych z rozkładu normalnego. Jeśli dane z niego nie pochodzą - wynik testu może być nieprawdziwy. Jednym ze sposobów zbadania tego faktu jest użycie testu χ^2 Pearsona. W celu jego przeprowadzenia należy wykonać następujące kroki:

1. utworzyć histogram z otrzymanych obserwacji i podzielić go na równe koszyki pamiętając żeby w każdym koszyku znalazło się najlepiej co najmniej 5 wyników,

2. zliczyć obserwacje w każdym koszyku n_i,

3. obliczyć $\overline{y}$ oraz u_y po wszystkich obserwacjach ze wzorów (1) i (2), a następnie przyjąć je za odpowiednio średnią oraz odchylenie standardowe dopasowywanego rozkładu Gaussa,

4. dla każdego koszyka policzyć wartość Np_i, gdzie p_i jest całką, dopasowywanego w poprzednim punkcie rozkładu, na odpowiadającym koszykowi przedziale,

5. obliczyć wartość statystyki χ^2 korzystając ze wzoru

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i},$$

6. porównać jej wartość z wartościami granicznymi dla rozkładu χ^2 o $i-3$ stopniach swobody.

Test można oczywiście również wykonać za pomocą różnych pakietów oprogramowania.

W przypadku gdy przyjęcie założenia o normalności rozkładu przeczy danym należałoby zamiast testu t Studenta zastosować jakiś test nieparametryczny istotności różnic średnich np. test Manna-Whitneya/Wilcoxona.

B. Metody nieparametryczne - test H Kruskala-Wallisa i test U Manna-Whitneya

TBD

III. METODA TAGUCHIEGO

Metoda Taguchiego w zastosowaniu do projektowania doświadczeń jest jednym ze sposobów na zredukowanie liczby wymaganych eksperymentów gdy jakaś wielkość zależy od k parametrów o $k_1, k_2, ..., k_k$ wartościach. Mniejsza liczba przeprowadzanych eksperymentów nie daje co prawda takiej dokładności jak w przypadku pełnego przetestowania wszystkich k^k możliwości, ale może być zupełnie wystarczająca w wielu przypadkach.

Jeśli poprawnie dobrano macierz Taguchiego oraz odpowiednio rozpisano różne kombinacje parametrów do doświadczeń, można wówczas użyć macierzy wyników do uporządkowania parametrów ze względu na ich wkład. Następujące kroki ilustrują przykładową analizę takiej macierzy.

1. obliczenie wszystkich $\overline{y_i}$ oraz u_i^2, gdzie te wielkości są liczone dla każdego eksperymentu i ze wszystkich powtórzeń (w tym wypadku będzie 16 par takich wielkości), używając wzorów (1) oraz (2),

2. obliczenie stosunku sygnału do szumu SN_i dla każdego eksperymentu w zależności od celu doświadczeń:

- w przypadku gdy chcemy maksymalizować obserwowany wynik

$$SN_i = -10\log\left[\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{1}{y_{i,j}^2}\right], \tag{4}$$

- w przypadku odwrotnym

$$SN_i = -10\log\left[\sum_{j=1}^{N_i}\frac{y_{i,j}^2}{N_i}\right],$$

3. obliczyć średnią arytmetyczną dla każdej wartości parametru $SN_{k,l}$, gdzie k jest numerem parametru, a l jedną z jego wartości. Takich średnich będzie oczywiście tyle ile suma wszystkich wartości przyjmowanych przez parametry, więc w przypadku 4ch parametrów z 4ma dopuszczalnymi wartościami każdy, będzie ich po prostu 16,

4. stworzyć macierz zawierającą wszystkie wartości $SN_{k,l}$ i dla każdego parametru k obliczyć wartość

$$R_k = \max(SN_{k,l}) - \min(SN_{k,l}).$$

Tak obliczone wielkości R_k są miarą wielkości wpływu odpowiednich parametrów na otrzymane wyniki. Znalezione wartości można następnie uporządkować malejąco/rosnącą w zależności od potrzeb w celu zilustrowania ich wkładów.

Analiza wariancji

Innym sposobem wykorzystania macierzy wyników jest analiza wariancji, a w naszym przypadku jej jednoczynnikowa wersja. Za hipotezę zerową w tej metodzie przyjmuje się równość średnich we wszystkich grupach, a założeniem przyjmowanym jest pochodzenie danych z rozkładu normalnego. Wykorzystuje się ją np. gdy chce się sprawdzić czy dany parametr ma istotny wpływ na otrzymywany wynik. Jeśli dla każdej dopuszczalnej wartości parametru otrzymalibyśmy te same średnie wartości wyniki to oznaczałoby, że nie ma on statystycznego wpływu na wynik. W związku z tym hipoteza zerowa dla 4ch różnych wartości parametru może przyjąć postać

$$\mu_0 : \overline{x_1} = \overline{x_2} = \overline{x_3} = \overline{x_4},$$

natomiast hipoteza przeciwna
μ_1 : nie wszystkie średnie są równe,
gdzie $\overline{x_i}$ oznacza średnią po wszystkich eksperymentach, dla których parametr k przyjmuje wartość k_i. W celu przeprowadzenia analizy wariancji należy:

1. podzielić wyniki na grupy ze względu na przyjmowane przez parametry wartości,

2. obliczyć sumę odchyleń po wszystkich wynikach dla wszystkich parametrów

$$\mathrm{SS_T} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{\overline{Y}})^2, \tag{5}$$

gdzie Y_{ij} to wynik eksperymentu dla parametru i o indeksie wartości j, a $\overline{\overline{Y}}$ to globalna średnia po wszystkich możliwych wynikach,

3. znaleźć sumy odchyleń średnich po konkretnych wartościach dla każdego z parametrów

$$\mathrm{SS_k} = \sum_{i=1}^{k}(\overline{Y_i} - \overline{\overline{Y}})^2, \tag{6}$$

gdzie $\overline{Y_i}$ jest średnią ze wszystkich obserwacji dla konkretnej wartości parametru k,

4. dla każdej wartości parametru zapisać wyrażenie $\mathrm{SS}_k/\mathrm{df}$, gdzie df jest liczbą stopni swobody (liczba wartości przyjmowanych przez dany parametr), analogicznie można podzielić $\mathrm{SS_T}$ przez sumę wszystkich stopni swobody,

5. ilorazy tak znalezionych wartości dla konkretnych parametrów przedstawiają udział danego parametru w objaśnianiu wyników.

Całość analizy można przedstawić w wyniku tabeli dla ANOVY[1].

IV. SZUKANIE NAJISTOTNIEJSZEGO PARAMETRU - PODEJŚCIE ALTERNATYWNE*

W przypadku gdy istnieje podejrzenie, że wynik eksperymentu jest liniową kombinacją jego parametrów, można zastosować metodę regresji liniowej. Jest to jedna z najprostszych metod uczenia maszynowego, która pozwala ilościowo wskazać, który parametr ma największy wpływ na końcowy wynik doświadczenia. Polega ona na takim przekształceniu parametrów x_i, aby ich posumowa z odpowiednimi współczynnikami wartość była najbliższa wynikowi eksperymentu. Poniższe równanie ilustruje powyższą metodę.

$$\boldsymbol{y} \sim f_1(\boldsymbol{x_1}) + f_2(\boldsymbol{x_2}) + f_3(\boldsymbol{x_3}) + f_4(\boldsymbol{x_4}),$$

gdzie $\boldsymbol{y}$ jest wektorem wyników, a kolejne $\boldsymbol{x_i}$, gdzie $i \in \{1, 2, 3, 4\}$, wektorami przekształconych parametrów. Zaletą tej metody jest jej stosunkowo łatwa interpretowalność oraz fakt, że łatwo stwierdzić, które parametry są istotne statystycznie w danym zagadnieniu. Głównym ograniczeniem takiej analizy jest wymóg wstępnego założenia o co najwyżej liniowym związku wkładów parametrów do wyniku, więc sytuacje, w których wynik otrzymywany jest poprzez jakieś przemnażanie parametrów przez

siebie lub bardziej złożone przekształcenia nie może być przez nią modelowany.

Istnieją również inne metody uczenia maszynowego, które są w stanie dać wgląd we względny wkład parametrów w otrzymaniu końcowego wyniku, niestety są one dużo bardziej złożone obliczeniowo, co raczej wykracza poza zakres analizy do tego typu pracy.

Źródła

1. Abdullah Naseer Mustapha, Yan Zhang, Zhibing Zhang, Yulong Ding, Qingchun Yuan, Yongliang Li, **Taguchi and ANOVA analysis for the optimization of the microencapsulation of a volatile phase change material**, Journal of Materials Research and Technology, Volume 11, 2021, Pages 667-680, ISSN 2238-7854.