ANOVA

ANOVA (ang. *ANalysis Of VAriance*) to zbiór metod i modeli statystycznych pozwalających na porównywanie dwu lub więcej populacji pod względem ich średniej. ANOVA jest szczególnie przydatna przy większej liczby grup niż 2, gdyż pozwala uniknąć – w przypadku korzystania z innych testów – wykonywania testów dla wszystkich możliwych par, tj. z pomocą metod z rodziny ANOVA porównujemy więcej grup jednocześnie, nie zaś wszystkie pary po kolei. W Matlabie metody z tej grupy są zaimplementowane jako część toolboxa Statistics (Statistics Toolbox > Regression and Classification > ANOVA).

UWAGA!
Jednoczesne porównywanie większej liczby grup NIE może być zastąpione wielokrotnym wykonywaniem testów dla porównania dwóch grup. Wynika to z konieczności kontrolowania błędu pierwszego rodzaju α. Wybierając poziom α i stosując k-krotnie wybrany test dla porównania dwóch grup zwiększylibyśmy znacznie założony poziom α. Tego błędu unikniemy stosując odpowiednie testy ANOVA.

Jednoczynnikowa analiza wariancji (one-way analysis of variance, one-way ANOVA)

Jednoczynnikowa analiza wariancji jest testem statystycznym dla danych numerycznych służącym do porównywania średnich w wielu populacjach. Zazwyczaj stosuje się ją do porównywania przynajmniej trzech grup jednocześnie, zaś przy dwu grupach pozostaje się przy t-teście.

Przyjmuje się następujące założenia:

- Populacje źródłowe muszą mieć rozkład normalny
- Próbki muszą być niezależne
- Wariancje poszczególnych populacji muszą być równe

przy czym dopuszcza się niewielkie odstępstwa od dwóch pierwszych założeń.

Testowana jest następująca hipoteza zerowa

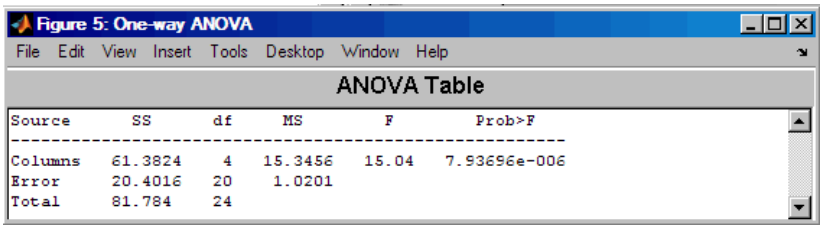
H_0: *próbki zawarte we wszystkich (dwu lub więcej) grupach pochodzą z populacji o jednakowych wartościach średnich*.
H_1: *dla przynajmniej jednej próbki średnia z niej jest znacząca inna od średnich z pozostałych próbek (*przekłada się to na średnie populacji*)*

Wykorzystanie funkcji *anova1* na macierzy X w Matlabie powoduje wyświetlenie dwu grup wyników. Po pierwsze, standardowej dla ANOVA tabeli oraz boxplotów wykonanych dla poszczególnych kolumn macierzy X. Standardowa tabela wyników rozróżnia dwa typy wariancji: wariancję wynikającą z różnic pomiędzy średnimi z kolumn (wariancję międzygrupową) oraz wariancję wynikającą z różnic pomiędzy danymi w poszczególnych kolumnach a ich odpowiadającymi średnimi (wariancję wewnątrzgrupową).

Standardowa tabela wyników dla ANOVA zawiera sześć kolumn:

- źródło wariancji (Source): międzygrupowa, wewnątrzgrupowa, łączna
- suma kwadratów (SS) odchyleń poszczególnych średnich od średniej globalnej

- liczba stopni swobody (df) dla każdego źródła
- MS (mean-square) wyliczone jako SS/df
- Statystykę F wyznaczoną jako MS(Columns)/MS(Error)
- P-value wyznaczone z wykorzystaniem dystrybuanty dla F



Przykład 1.

Utwórz następujące dane i uruchom dla nich metodę anova1. Następnie przeanalizuj wyniki.

```
X = meshgrid(1:5);
X = X + normrnd(0,1,5,5);
p = anova1(X);
```

Uwaga: Jeżeli dane w poszczególnych grupach nie są równoliczne, konieczne jest ich wcześniejsze przygotowanie.

Przykład 2.

Załóżmy, że mamy dane zawarte w jednym wektorze o 20 elementach:

X = [82 86 79 83 84 85 86 87 74 82 ...
78 75 76 77 79 79 77 78 82 79];

Niech pierwsze 8 wyników pochodzi z testów wytryzmałościowych wykonanych na stali (steel), kolejne 6 z testów na pierwszym stopie (alloy1) i ostatnie 6 z testów na drugim stopie (alloy2). Należy wtedy przygotować dodatkową zmienną która opisze grupy danych. Może ona mieć postać zmiennej typu cell o jednakowej długości zawierającej identyfikatory grup dla kolejnych danych z pierwotnego wektora, np:

```
group = {'st','st','st','st','st','st','st','st',...
         'al1','al1','al1','al1','al1','al1',...
         'al2','al2','al2','al2','al2','al2'};
```

Możliwe są też inne reprezentacje grup (Matlab: Grouping Variables). Na przykład za pomocą zwykłego wektora liczb. Dla powyższego przykładu mógłby on mieć postać:

```
group = [1,1,1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3];
```

Następnie wywołujemy metodę anova1 korzystając z wektora z danymi i z wektora opisującego przynależność grupową danych: p = anova1(X,group).

Ćwiczenie 1: Wczytaj dane z pliku anova_data dołączone do listy zadań. Dane w zmiennej koala przedstawiają dobowy czas snu trzech grup zwierząt. Zweryfikować czy dane w poszczególnych grupach pochodzą z niezależnych populacji o rozkładzie normalnym i o równych wariancjach (potraktować to jako wprawkę do końcowych list: wybrać test, sformułować hipotezy, zinterpretować wynik). Jeżeli tak, wykorzystać jednoczynnikową analizę wariancji w celu zweryfikowania odpowiedniej hipotezy dotyczącej równości średnich w powyższych trzech populacjach. Zinterpretować wynik.

Ćwiczenie 2 Dane w zmiennych wombats i wombats_groups przedstawiają dane dotyczące aktywności dobowej wombatów. Przyjmując, że dane w poszczególnych grupach pochodzą z niezależnych populacji o rozkładzie normalnym i o równych wariancjach wykorzystać jednoczynnikową analizę wariancji w celu zweryfikowania odpowiedniej hipotezy dotyczącej równości średnich we wszystkich populacjach. Zinterpretować wynik.

Warto zwrócić uwagę, że jednoczynnikowa analiza wariancji bada wpływ jednej zmiennej niezależnej na zmienną zależną. Analiza średniej jest wystarczająca, jako że założenia o danych są już bardzo daleko idące: m.in. rozkłady normalne w populacjach i jednakowe wariancje.

Jeśli wynik tego testu pozwala nam odrzucić hipotezę zerową o równości średnich to rodzi się pytanie które średnie są różne. Jest wiele testów które na tym etapie można wykonać aby odpowiedzieć na to pytanie. W matlabie jest zaimplementowany test oparty na badaniu przedziałów ufności dla średnich

```
[c,m]=multcompare(stats)
```

stats jest macierzą zwracaną przez anova1. Macierz c zawiera następujące informacje: numer pierwszej grupy, numer drugiej grupy, dolny brzeg przedziału ufności (95%), różnica między średnimi tych dwóch grup, górny brzeg przedziału ufności, p-value, dla hipotezy zerowej, że średnie w tych grupach są sobie równe. Te informacje są także prezentowane w oknie graficznym otwieranym przez funkcję multcompare.

Jednoczynnikowa analiza wariancji – Kruskal-Wallis oraz Friedman

Jeżeli założenia dla ANOVY (szczególnie normalność rozkładu i równość wariancji) nie są spełnione stosujemy test nieparametryczny Kruskala-Wallisa albo Friedmana. Testy te służą do weryfikacji hipotezy o nieistotności różnic pomiędzy medianami badanej zmiennej w kilku (k >= 2)populacjach (przy czym zakładamy, że rozkłady zmiennej są sobie bliskie).

Podstawowe warunki stosowania:
− pomiar na skali porządkowej lub interwałowej,
− model niezależny.

```
[p,table] = kruskalwallis(X,group)
p = friedman(X,group)
```

Ćwiczenie 3 Badano kwartalną wielkość sprzedaży pewnego batonu czekoladowego w 14 losowo wybranych marketach. Badanie rozpoczęto w styczniu a zakończono w grudniu. W czasie drugiego kwartału trwała intensywna billboardowa kampania reklamowa tego produktu. Sprawdzić, czy kampania miała wpływ na wielkość sprzedaży reklamowanego batonu.

sklep	Kwartał I	Kwartał II	Kwartał III	Kwartał IV
Sk1	3415	4556	5772	5432

Sk2	1593	1937	2242	2794
Sk3	1976	2056	2240	2085
Sk4	1526	1594	1644	1705
Sk5	1538	1634	1866	1769
Sk6	983	1086	1135	1177
Sk7	1050	1209	1245	977
Sk8	1861	2087	2054	2018
Sk9	1714	2415	2361	2424
Sk10	1320	1621	1624	1551
Sk11	1276	1377	1522	1412
Sk12	1263	1279	1350	1490
Sk13	1271	1417	1583	1513
Sk14	1436	1310	1357	1468

Dwuczynnikowa analiza wariancji (Two-way analysis of variance, two-way ANOVA)

W odróżnieniu od jednoczynnikowej analizy wariancji, dwuczynnikowa analiza wariancji jest testem statystycznym dla danych numerycznych służącym do porównywania rozkładów pewnej wartości w grupach opisanych dwoma czynnikami (dwiema zmiennymi niezależnymi). Można na przykład badać stężenie ołowiu we krwi w grupach pracowników z dwu różnych manufaktur pracujących z wykorzystaniem dwu różnych technologii w każdej z manufaktur; w tej sytuacji byłyby cztery grupy: manufaktura1-technologia1, manufaktura1-technologia2, manufaktura2-technologia1 i manufaktura2-technologia2. Porównanie grup odbywa się poprzez odwołanie do równości średnich w poszczególnych populacjach.

Przyjmuje się następujące założenia:

- Populacje źródłowe muszą mieć rozkład normalny
- Próbki muszą być niezależne
- Wariancje poszczególnych populacji muszą być równe
- Grupy muszą być równoliczne

przy czym dopuszcza się niewielkie odstępstwa od dwóch pierwszych założeń.

Testowana są następujące hipotezy zerowe

H_{01}: *średnie populacyjne wyznaczone względem pierwszego czynnika są równe (one-way ANOVA dla kolumn jako populacji)*

H_{02}: *średnie populacyjne wyznaczone względem drugiego czynnika są równe (one-way ANOVA dla grup wierszy)*

H_{03}: *oba czynniki nie mają synergicznego wpływu na średnie populacyjne (tj. kombinacja czynników nie ma dodatkowego wpływu na średnie populacyjne – w porównaniu z wpływem z osobna)*

Przykład 3.

Wykorzystamy dane przechowywane w zmiennej popcorn.

```
popcorn =
  5.5000  4.5000  3.5000
```

```
5.5000   4.5000   4.0000
6.0000   4.0000   3.0000
6.5000   5.0000   4.0000
7.0000   5.5000   5.0000
7.0000   5.0000   4.5000
```

Wiemy, że w kolumnach przedstawione są wyniki dotyczące popcornu pochodzącego od trzech różnych producentów. Pierwsze trzy wiersze dotyczą wyników uzyskanych w przypadku zastosowania maszyny powietrznej, a drugie trzy wiersze w przypadku zastosowania maszyny olejowej. Mierzone wartości przedstawiają ilość kubków popcornu uzyskanego z jednej miarki.

Z niezależnych pomiarów wiemy, że liczba kubków w każdym z przypadków ma rozkład normalny o jednakowej wariancji. Wiemy też, że pomiary dokonane w poszczególnych populacjach są niezależne. Ponieważ ponadto grupy są równoliczne (na każdej populacji dokonano trzech pomiarów), można wykorzystać dwuczynnikową analizę wariancji w celu weryfikacji wpływu poszczególnych czynników na uzyskiwaną ilość popcornu.

Stawiamy następujące hipotezy:

H_{01}: *średnia liczba kubków dla każdego z producentów jest jednakowa*
H_{02}: *średnia liczba kubków jest niezależna od typu maszyny*
H_{03}: *producent i typ maszyny nie mają synergicznego wpływu na średnie populacyjne*

Uruchamiamy metodę dla powyższych danych jako drugi argument podając liczbę wierszy, które zajmuje każda z populacji:

```
p = anova2(popcorn,3)
p = 0.0000   0.0001   0.7462
```

Uzyskane wartości p wskazują, że są podstawy do odrzucenia hipotez zerowych **H_{01}** i **H_{02}** (o braku wpływu producenta i typu maszyny na ilość otrzymywanego produktu). Jednocześnie przy poziomie istotności 0.05 nie ma podstaw do odrzucenia hipotezy zerowej **H_{03}** mówiącej, że producent i typ maszyny nie mają dodatkowego synergicznego wpływu na średnią ilość otrzymywanego produktu.

Poszczególne wartości w zamieszczonej poniżej tabeli wyników mają analogiczną interpretację, jak w jednoczynnikowej analize wariancji. Dodatkowym czynnikiem są wartości wyznaczone w wierszu *Interaction* i większa liczba weryfikowanych hipotez.

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	15.75	2	7.875	56.7	0
Rows	4.5	1	4.5	32.4	0.0001
Interaction	0.0833	2	0.04167	0.3	0.7462
Error	1.6667	12	0.13889		
Total	22	17			

Ćwiczenie 4 Dla danych z przykładu 3 wykonaj analizę post-hoc wykorzystując do tego test multicompare. Przeanalizuj otrzymane wyniki, wyciągnij odpowiednie wnioski.

Ćwiczenie 5 Badano wpływ trzech różnych substancji toksycznych (ozn.: T1, T2, T3) stosowanych w procesie produkcyjnym na układ oddechowy pracowników trzech różnych

zakładów przemysłowych (ozn.: Z1, Z2, Z3). Jako miarę wydolności oddechowej przyjęto objętość wymuszonego wydechu FED (ang. *forced expiratory volume*). Zebrane dane dla 108 pracowników przedstawia tabela.

FEV								
T1			T2			T3		
Z1	Z2	Z3	Z1	Z2	Z3	Z1	Z2	Z3
4,64	5,12	4,64	3,21	3,92	4,95	3,75	2,95	2,95
5,92	6,10	4,32	3,17	3,75	5,22	2,50	3,21	2,80
5,25	4,85	4,13	3,88	4,01	5,16	2,65	3,15	3,63
6,17	4,72	5,17	3,50	4,64	5,35	2,84	3,25	3,85
4,20	5,36	3,77	2,47	3,63	4,35	3,09	2,30	2,19
5,90	5,41	3,85	4,12	3,46	4,89	2,90	2,76	3,32
5,07	5,31	4,12	3,51	4,01	5,61	2,62	3,01	2,68
4,13	4,78	5,07	3,85	3,39	4,98	2,75	2,31	3,35
4,07	5,08	3,25	4,22	3,78	5,77	3,10	2,50	3,12
5,30	4,97	3,49	3,07	3,51	5,23	1,99	2,02	4,11
4,37	5,85	3,65	3,62	3,19	4,76	2,42	2,64	2,90
3,76	5,26	4,10	2,95	4,04	5,15	2,37	2,27	2,75

Należy dostosować dane do wykorzystania w Matlabie. Następnie: sprawdzić założenia i jeżeli to możliwe, zastosować dwuczynnikową analizę wariancji (sformułowanie hipotez, przeprowadzenie testu, analiza wyników).

Istnieje analogiczna metoda dla analizy wpływu większej liczby zmiennych niezależnych (anovan). Nie będzie ona tutaj analizowana, ale można się z nią zapoznać na stronie dokumentacji Mathworks.