Analiza Wariancji (ANOVA)

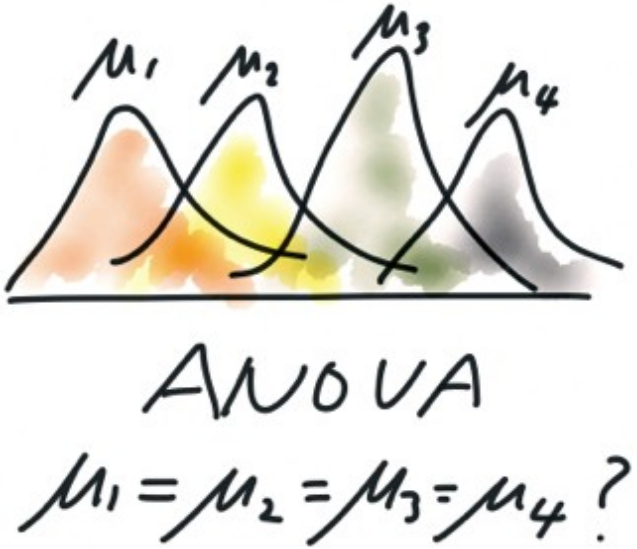
Statystyka biomedyczna

Piotr M. Szczypiński

Czym jest ANOVA

Analiza wariancji ANOVA to w rzeczywistości grupa analiz statystycznych, służących do badania wpływu czynników (zmiennych niezależnych) na zmienną zależną. ANOVA jest popularna i często stosowana (nawet nadużywana) ze względu na jej “elagancję” (związek z metodą najmniejszych kwadratów) i prostotę obliczeniową.

ANOVA umożliwia przeprowadzenie testu statystycznego dla większej liczby grup niż dwie, czym odróżnia się od t-testu Studenta.



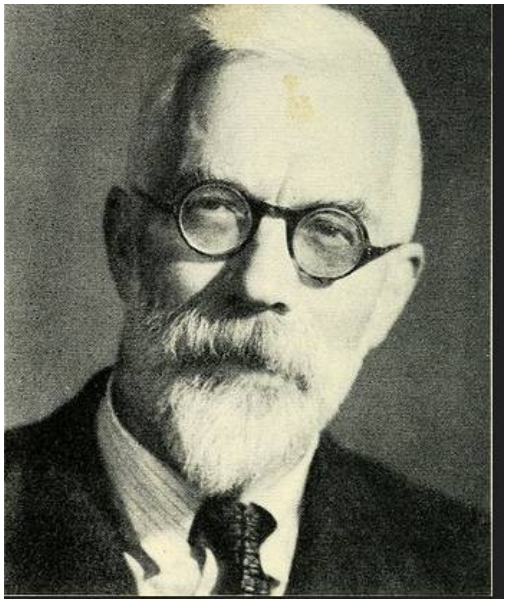
https://www.numberanalytics.com/anova_upload.php

https://en.wikipedia.org/wiki/Analysis_of_variance
http://www.naukowiec.org/wiedza/statystyka/analiza-wariancji-anova_738.html

Czym jest ANOVA

ANOVA została zaproponowana i opracowana w latach dwudziestych XX wieku przez Sir Ronalda Aylmera Fishera, angielskiego biologa i statystyka.

To on zaproponował użycie określenia wariancja. W analizie wariancji wykorzystane zostały wcześniejsze osiągnięcia Laplacea i Gaussa (metoda najmniejszych kwadratów) oraz Jerzego Neymana (modele randomizacji)



Sir Ronald Fisher
(1890 – 1962)

https://en.wikipedia.org/wiki/Analysis_of_variance
https://en.wikipedia.org/wiki/Ronald_Fisher

Modele analizy wariancji

modele jednoczynnikowe – badany jest wpływ pojedynczego czynnika lub wpływ każdego czynnika jest rozpatrywany oddzielnie. Jest to jednoczynnikowa analiza wariancji (ang. one-way ANOVA),

modele wieloczynnikowe - wpływ różnych czynników jest rozpatrywany łącznie. Tą klasą zagadnień zajmuje się wieloczynnikowa analiza wariancji (ang. multi-way ANOVA).

Zastosowanie - przykład

http://mazurytravel.com.pl/wydarzenia/xvi-wystawa-psow-rasowych

Psy na wystawie:

- nie są losową próbą całej populacji (dorosłe, wyselekcjonowane)
- psy są róznych ras, wielkości i o różnej wadze.

Cel:

- Przewidywanie wagi psa na podstawie jego wybranych charakterystyk jakościowych

Na podstawie:
https://en.wikipedia.org/wiki/Analysis_of_variance

Zastosowanie - przykład

Proponujemy wstępnie podział na grupy zgodnie z dwoma czynnikami, wiekiem i długością sierści:

- starsze i młodsze
- krótko- i długowłose

- Y: wszystkie psy z wystawy
- X1: młodsze i krótkowłose
- X2: młodsze i długowłose
- X3: starsze i krótkowłose
- X4: starsze i długowłose

Na podstawie wykresu:
brak istotnej zależności (H0)

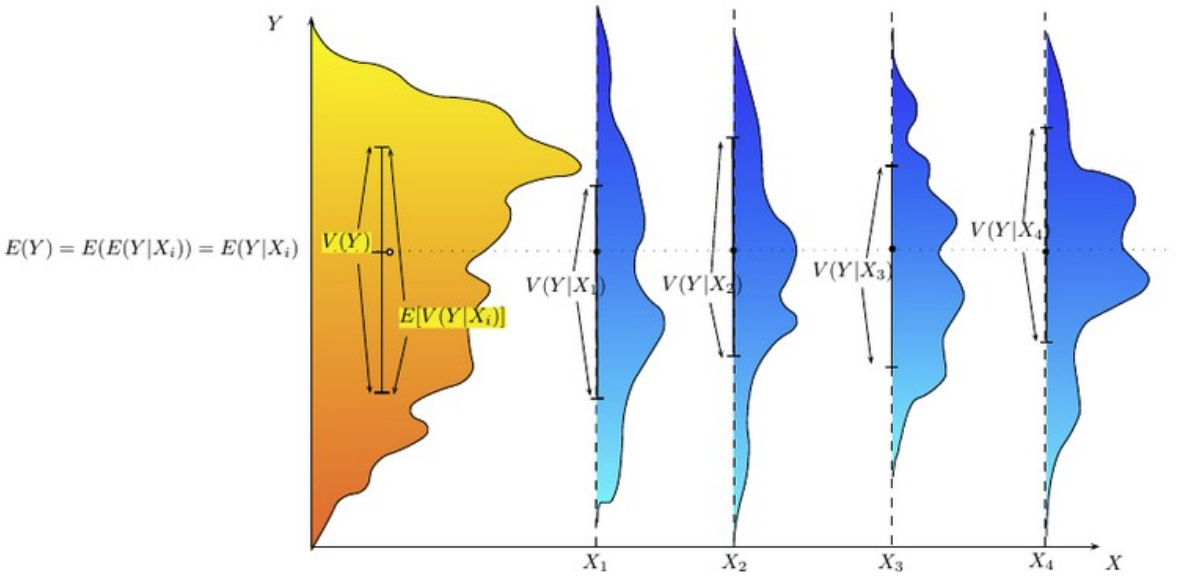


Figure 2: ANOVA : No fit

Zastosowanie – przykład c.d.

Proponujemy kolejny podział na grupy zgodnie z dwoma innymi czynnikami:

- hodowane do pracy oraz domowe
- mniej i bardziej atletycznej budowy

- Y: wszystkie psy z wystawy
- X1: domowe, mniej atletyczne
- X2: domowe, bardziej atletyczne
- X3: do pracy, mniej atletyczne
- X2: do pracy, bardziej atletyczne

Na podstawie wykresu:
widoczna **niewielka zależność (H0 ? H1)**

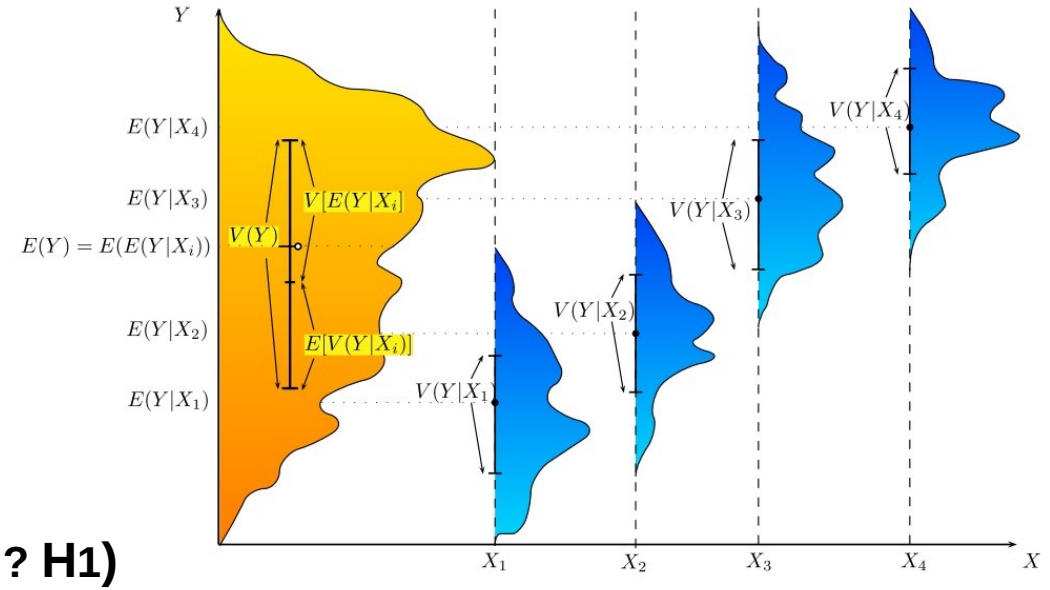


Figure 1: ANOVA : Fair fit

Zastosowanie – przykład c.d.

Proponujemy kolejny podział, w którym uwzględnimy jeden czynnik – rasę.

- Y: wszystkie psy z wystawy
- X1: Pudelek
- X2: Wyżeł niemiecki
- X3: Rottweiler
- X2: Bernardyn

Na podstawie wykresu:
wyraźna zależność (H1)

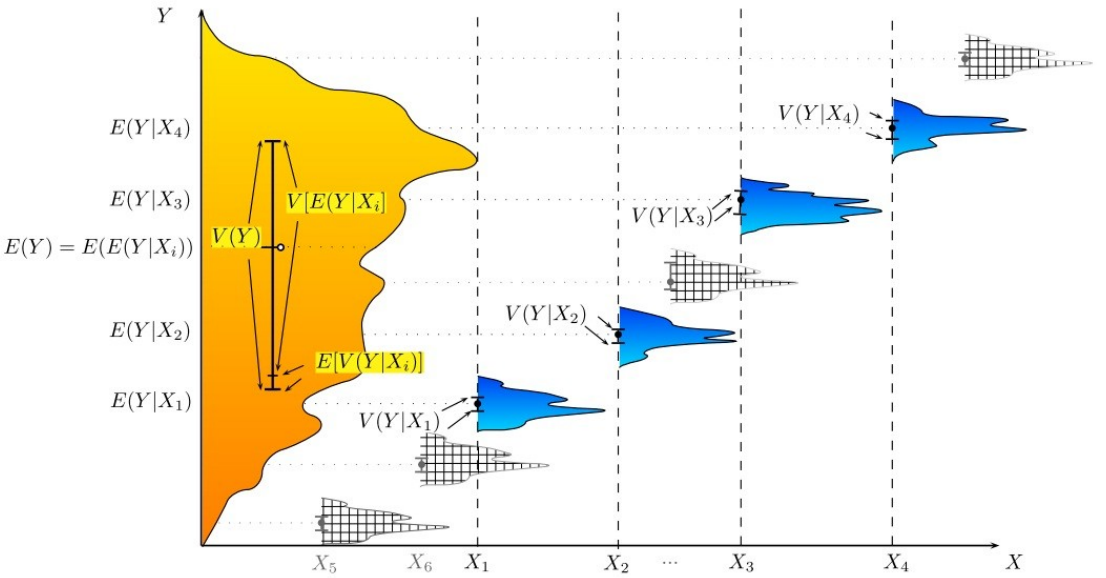


Figure 3: ANOVA : very good fit

Założenia

Wyniki uzyskane metodą analizy wariancji mogą być uznane za prawdziwe, gdy spełnione są następujące założenia:

- Populacje muszą mieć rozkłady normalne
- Próby są niezależne
- Próby każdej populacji muszą być losowe
- Wariancje w populacjach są równe

Gdy założenia analizy wariancji nie są spełnione należy użyć testu Kruskala-Wallisa.

https://pl.wikipedia.org/wiki/Jednoczynnikowa_analiza_wariancji

Hipotezy

Hipoteza zerowa:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_r$$

Hipoteza alternatywna:

$$H_1 : nie\ wszystkie\ \mu_i\ są\ sobie\ równe\ (i = 1, 2, ... r)$$

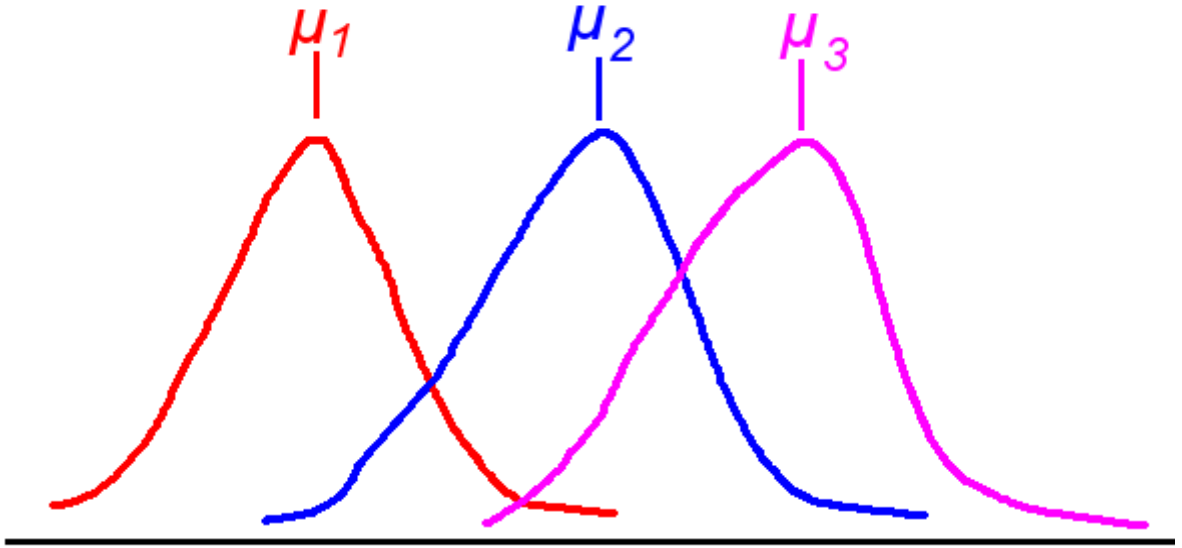
Jednoczynnikowa analiza wariancji

Statystyka *F*:

$$F = \frac{\frac{1}{r-1}\sum_{i=1}^{r} n_i(\mu_i - \mu)^2}{\frac{1}{n-r}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \mu_i)^2}$$

$$\mu_i = \frac{1}{r}\sum_{j=1}^{n_i} x_{ij}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}$$



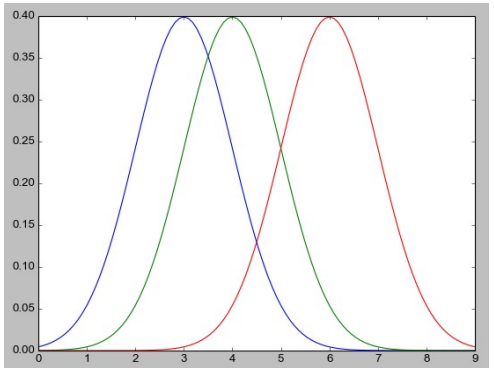
r – liczba populacji
n_i – wielkość populacji *i*
n – wielkość całej próby
$n = n_1 + n_2 + \ldots n_r$

Przykład

2.7529 3.4175 2.5336 2.0586 2.3763 1.4655 4.9820 3.8539

2.9074 5.5708 4.0635 2.8651 3.7384 3.2082 5.7290 1.8234

5.0922 7.6382 5.7826 5.8355 6.6824 7.5945 5.8338 5.0653



Uwaga: w tej prezentacji pomijamy sprawdzenie (testy) czy spełnione są założenia (rozkłady są normalne o równych wariancjach).

Przykład

2.7529	3.4175	2.5336	2.0586	2.3763	1.4655	4.9820	3.8539	$\mu_1 = 2.9300$
2.9074	5.5708	4.0635	2.8651	3.7384	3.2082	5.7290	1.8234	$\mu_2 = 3.7382$
5.0922	7.6382	5.7826	5.8355	6.6824	7.5945	5.8338	5.0653	$\mu_3 = 6.1905$

$\mu = 4.2863$

$n_1 = n_2 = n_3 = 8$

$n = n_1 + n_2 + n_3 = 24$

$r = 3$

$$F = \frac{\frac{1}{r-1}\sum_{i=1}^{r} n_i(\mu_i - \mu)^2}{\frac{1}{n-r}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \mu_i)^2} = \frac{\frac{1}{2}\left(8(2.9300-4.2863)^2 + 8(3.7382-4.2863)^2 + 8(6.1905-4.2863)^2\right)}{\frac{1}{21}(1.2876+1.4395+0.9590)} = ?$$

p-wartość należy odczytać z tablic rozkładu F Snedecora dla dwóch stopni swobody r-1 = 2 i n-r = 21.

Przykład

W Pythonie do obliczenia testu służy f_oneway()

```
>>> y1
array([ 2.7529, 3.4175, 2.5336, 2.0586, 2.3763, 1.4655, 4.9820, 3.8539])
>>> y2
Array([ 2.9074, 5.5708, 4.0635, 2.8651, 3.7384, 3.2082, 5.7290, 1.8233])
>>> y3
array([ 5.0922, 7.6382, 5.7826, 5.8355, 6.6824, 7.5945, 5.8338, 5.0653])
>>> f_oneway(y1, y2, y3)
(16.858883019566143, 4.2949175811022016e-05)
```

F = 16.86 *p*-value = 0.000043 < 0.01

Materiały w sieci

James Jones, Statistics: Lecture Notes, Chapter 13

https://people.richland.edu/james/lecture/m170/

Wideo prezentacje

Wprowadzenie ANOVA

- https://www.youtube.com/watch?v=0Vj2V2qRU10

One-way (jednoczynnikowa) ANOVA

- https://www.youtube.com/watch?v=JgMFhKi6f6Y
- https://www.youtube.com/watch?v=UrRYITjDOww

Two-way (dwuczynnikowa) ANOVA

- https://www.youtube.com/watch?v=i4mamuL_CF8
- https://www.youtube.com/watch?v=qdPJRP3j5WM