Statystyka

Martyna Śpiewak

Bootcamp Data Science

Testy zgodności

Załóżmy, że interesuje nas **postać rozkładu** badanej cechy X. Dystrybuantę owego nieznanego rozkładu oznaczymy przez F.

Hipoteza dotycząca F może być dwojakiego rodzaju

- **hipoteza prosta**

  $$H_0 : F = F_0$$

  np. X ma rozkład wykładniczy o wartości oczekiwanej 100;
- **hipoteza złożona**

  $$H_0 : F \in \mathcal{F},$$

  gdzie $\mathcal{F}$ oznacza pewną rodzinę dystrybuant (rozkładów), np. X ma rozkład wykładniczy.

Test zgodności chi-kwadrat

Założenia:

- duża próba, $n \geq 100$;
- rozkład dyskretny lub ciągły;
- dane pogrupowane w szereg rozdzielczy, w taki sposób, aby liczności poszczególnych klas nie były mniejsze niż 5 ($n_i \geq 5$).

Hipotezę zerową

$$H_0 : F = F_0,$$

w której dystrybuanta F_0 jest w pełni określona, testujemy porównując zaobserwowane liczności n_i z licznościami hipotetycznymi np_i odpowiadającymi oczekiwanym licznościom poszczególnych klas przy założeniu hipotezy zerowej.

Test zgodności chi-kwadrat

Wielkości p_i obliczamy ze wzoru

$$p_i = F_0(\xi_i) - F_0(\xi_{i-1}),$$

gdzie $\xi_0, \dots, \xi_k$ oznaczają krańce przedziałów klasowych.

Statystyka testowa

$$T = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

ma przy prawdziwości H_0 asymptotyczny rozkład chi-kwadrat o $k-1$ stopniach swobody. Obszar krytyczny testu na przyjętym poziomie istotności α będzie miał postać

$$W_\alpha = [\chi^2_{1-\alpha,k-1}, +\infty],$$

gdzie $\chi^2_{1-\alpha,k-1}$ oznacza kwantyl rzędu $1-\alpha$ rozkładu chi-kwadrat o $k-1$ stopniach swobody.

Test zgodności chi-kwadrat — przykład

Znaleźliśmy złotego dukata, który z jednej strony ma wizerunek cesarza, a z drugiej akwedukt. Chcemy stwierdzić czy moneta jest sprawiedliwa czyli, czy rzucając nią mam równą szansę na wypadnięcie akweduktu lub cesarza. W tym celu wykonuję 100 rzutów i otrzymuję następujące rezultaty:

Cesarz	Akwedukt
63	37

Test zgodności chi-kwadrat — przykład

- X - zmienna losowa przyjmująca wartości 1 gdy wypada cesarz i 0 gdy wypada akwedukt.
- Badamy następującą parę hipotez:

$$H_0 : X \sim Bern(0.5)$$

  $\mathrm{H}_1 : \neg H_0$

- $n_1 = 63$, $n_2 = 37$
- $p_1 = p_2 = 0.5$
- $np_1 = np_2 = 50$
- $T = \frac{(63-50)^2}{50} + \frac{(37-50)^2}{50} = \frac{2*13^2}{50} = 6.76$
- Przedział krytyczny $[\chi^2_{0.95,1}, \infty] \approx [3.84, \infty]$
- p-value ≈ 0.009.
- odrzucamy H_0 na rzecz H_1.

Test zgodności chi-kwadrat

W praktycznych zastosowaniach na ogół nie znamy rozkładu F_0. Wówczas testujemy hipotezę złożoną, w której $\mathcal{F}$ jest pewną rodziną rozkładów zależną od r nieznanych parametrów $\theta_1, \ldots, \theta_r$, tzn.

$$H_0 : F \in \mathcal{F} = \{F : F = F_{\theta_1,\ldots,\theta_r}\}.$$

W tym przypadku

1. szacujemy nieznane parametry (najlepiej metodą największej wiarogodności) otrzymując $\hat{\theta}_1, \ldots \hat{\theta}_r$;
2. rozważaną hipotezę złożoną zastępujemy hipotezą prostą

$$H_0 : F = F_{\hat{\theta}_1,\ldots,\hat{\theta}_r}.$$

Test zgodności chi-kwadrat

Statystyka testowa pozostaje bez zmian, tyle, że wielkości p_i obliczamy ze wzoru

$$p_i = F_{\hat{\theta}_1,\ldots,\hat{\theta}_r}(\xi_i) - F_{\hat{\theta}_1,\ldots,\hat{\theta}_r}(\xi_{i-1}).$$

Przy prawdziwości hipotezy H_0 statystyka ta ma asymptotyczny rozkład chi-kwadrat o $k-r-1$ stopniach swobody, gdzie r jest liczbą estymowanych parametrów, a obszar krytyczny testu na przyjętym poziomie istotności α ma postać

$$W_\alpha = [\chi^2_{1-\alpha,k-r-1}, +\infty],$$

Test Kołmogorowa

Istotą testu jest porównanie dystrybuanty empirycznej $\hat{F}_n$ zbudowanej na podstawie próby, z dystrybuantą teoretyczną F_0.

Założenia:

- dowolna liczność próbki;
- rozkład próby ciągły.

Za statystykę testową do weryfikacji hipotezy prostej Kołmogorow przyjął miarę odległości dystrybuanty empirycznej od dystrybuanty teoretycznej

$$D_n = \sup_x |\hat{F}_n(x) - F_0(x)|.$$

Przy założeniu prawdziwości hipotezy H_0 statystyka D_n ma rozkład niezależny od F_0. Wartości krytyczne $D_n(\alpha)$ są stablicowane. Obszar krytyczny testu ma postać

$$W_\alpha = [D_n(\alpha), 1].$$

Test Kołmogorowa-Smirnowa

Test Kołmogorowa-Smirnowa stosowany jest do weryfikacji hipotezy

$$H_0 : F_1 = F_2,$$

o identyczności rozkładów badanej cechy dla dwóch populacji, wobec hipotezy alternatywnej orzekającej, że rozkłady te istotnie się różnią.

Założenie: Rozkłady badanych cech powinny być ciągłe.

Niech $X_1, \ldots, X_n$ będzie próbą losową pochodzącą z pierwszej populacji, natomiast $Y_1, \ldots, Y_m$ próbą losową pochodzącą z drugiej populacji.

Statystyką testową jest

$$D_{n,m} = \sup_x |\hat{F}_n(x) - \hat{F}_m(x)|,$$

gdzie $\hat{F}_n(x)$ i $\hat{F}_m(x)$ oznaczają, odpowiednio, dystrybuanty empiryczne wyznaczone na podstawie pierwszej i drugiej próbki.

Zbyt duże wartości tej statystyki świadczą przeciw hipotezie zerowej, stąd obszar krytyczny testu ma postać

$$W_\alpha = [d(\alpha, n, m), 1],$$

gdzie $d(\alpha, n, m)$ jest wartością krytyczną rozkładu statystyki $D_{n,m}$.

Test Kruskala-Wallisa

Uogólnienie rozważanego przypadku testowania hipotezy o identyczności rozkładów badanej cechy dla dwóch populacji jest weryfikacja hipotezy o identyczności rozkładów dla k populacji, gdzie $k > 2$, tzn.

$$H_0 : F_1 = F_2 = \ldots = F_k,$$

wobec hipotezy alternatywnej, że rozkład badanej cechy nie we wszystkich populacjach jest taki sam.

Założenia: Rozważane rozkłady powinny być ciągłe.

Test Kruskala-Wallisa — algorytm

1. Załóżmy, że mamy k próbek o licznościach $n_1, \ldots, n_k$, przy czym $\sum_{i=1}^{k} n_i = n$;
2. Obserwacje pochodzące ze wszystkich k prób ustawiamy w porządku rosnącym;
3. Numerujemy kolejnymi liczbami naturalnymi (nadajemy tzw. *rangi*). Jeżeli kilka kolejnych wyników ma tę samą wartość, to każdemu z nich przypisujemy rangę będącą średnią arytmetyczną przypisanych im liczb naturalnych;
4. Dla każdej próbki oddzielnie wyznaczamy sumę rang R_i, po czym obliczamy wartość statystyki testowej.

Test Kruskala-Wallisa — statystyka testowa

Postać statystyki testowej testu Kruskala-Wallisa jest postaci

$$T = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\frac{R_i}{n_i} - \frac{(n+1)}{2} \right)^2$$

$$= \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1).$$

Przy założeniu prawdziwości hipotezy zerowej statystyka ma asymptotyczny rozkład chi-kwadrat o $k-1$ stopniach swobody.

Obszar krytyczny ma postać

$$W_\alpha = [\chi^2_{1-\alpha,k-1}, +\infty),$$

gdzie $\chi^2_{1-\alpha,k-1}$ oznacza kwantyl rzędu $1-\alpha$ rozkładu chi-kwadrat o $k-1$ stopniach swobody (tj. *duże wartości statystyki świadczą przeciwko hipotezie zerowej*).

Niejednokrotnie badając pewną populację mamy informację dotyczące dwóch cech owej populacji i w związku z tym interesuje nas, czy owe cechy są niezależne, czy też występuje między nimi jakaś zależność.

Rozważmy próbę, która jest ciągiem par

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

gdzie X_i oraz Y_i oznaczają, odpowiednio, wartości pierwszej i drugiej cechy przyjmowane przez i-ty element próby.

Tablica korelacyjna (dwudzielcza)

Formą zapisu wyników badania będącego ciągiem par jest **tablica korelacyjna**.

Tablica korelacyjna:

- zawiera tyle wierszy, ile wyróżniamy poziomów pierwszej cechy (załóżmy, że wyróżniliśmy r różnych poziomów pierwszej cechy);
- zawiera tyle kolumn, ile wyróżniamy poziomów drugiej cechy (załóżmy, że wyróżniliśmy c różnych poziomów drugiej cechy);
- tabela korelacyjna ma wymiar $r \times c$;
- wewnątrz każdej komórki tabeli wpisuje się liczbę tych elementów próby, dla których zaobserwowano wartości poziomu pierwszej cechy odpowiadający poziomowi tego wiersza i jednocześnie wartość drugiej cechy równą poziomowi danej kolumny.

Test niezależności chi-kwadrat

Weryfikować będziemy hipotezę

$$H_0 : \text{ cechy są niezależne}$$

wobec hipotezy alternatywnej

$$H_1 : \text{ cechy są zależne.}$$

Statystyka testowa jest dana wzorem

$$T = \sum_{j=1}^{rc} \frac{(O_j - E_j)^2}{E_j},$$

gdzie O_j oznacza liczbę obserwacji w j-tej komórce tabeli korelacyjnej, natomiast E_j jest tzw. *oczekiwaną liczbą obserwacji*, która powinna znaleźć się w j-tej komórce, jeżeli rozpatrywanej cechy są istotnie niezależne.

Test niezależności chi-kwadrat

Oczekiwaną liczbę obserwacji wylicza się dla każdej komórki ze wzoru

$$E_j = \frac{\sum_j^r \sum_j^c}{n},$$

gdzie $\sum_j^r$ oznacza sumę obserwacji w wierszu, w którym położona jest j-ta komórka, $\sum_j^c$ jest sumą obserwacji w kolumnie, do której należy j-ta komórka, zaś n jest licznością próby.

Przy założeniu prawdziwości hipotezy zerowej oraz dla licznej próby $(n \geq 100)$, statystyka testowa T ma w przybliżeniu rozkład chi-kwadrat o $(r-1)(c-1)$ stopniach swobody.

Duże wartości statystyki testowej przemawiają przeciwko hipotezie zerowej. Stąd obszar krytyczny ma postać

$$W_\alpha = [\chi^2_{(1-\alpha,(r-1)(c-1)}, +\infty).$$

Test niezależności — przykład

H_0 : nie ma związku między dochodem a preferencjami politycznymi,

H_1 : $\neg H_0$.

		Dochody		
		wysokie	**niskie**	$\sum$
Preferencje polityczne	**D**	60	110	170
	R	75	55	130
	$\sum$	135	165	300

Test niezależności — przykład

i	O_k	E_k	$\frac{(O_k-E_k)^2}{E_k}$
1.	60	76.5	3.56
2.	110	93.5	2.91
3.	75	58.5	4.65
4.	55	71.5	3.81
$\sum$	300	300	$t = 14.98$

1. Wyznaczamy obszar krytyczny:
   $W_{0.05} = [\chi^2_{0.95,1}, +\infty) = [3.84, +\infty) \implies t \in W_{0.05}$.

Test niezależności — przykład

i	O_k	E_k	$\frac{(O_k - E_k)^2}{E_k}$
1.	60	76.5	3.56
2.	110	93.5	2.91
3.	75	58.5	4.65
4.	55	71.5	3.81
$\sum$	300	300	$t = 14.98$

1. Wyznaczamy obszar krytyczny:

   $W_{0.05} = [\chi^2_{0.95,1}, +\infty) = [3.84, +\infty) \implies t \in W_{0.05}$.
2. Wyznaczamy p-wartość: $P(T > t | H_0) = 1 - F_{\chi^2_1}(t) = 0.0001$.

Test niezależności — przykład

i	O_k	E_k	$\frac{(O_k-E_k)^2}{E_k}$
1.	60	76.5	3.56
2.	110	93.5	2.91
3.	75	58.5	4.65
4.	55	71.5	3.81
$\sum$	300	300	$t = 14.98$

1. Wyznaczamy obszar krytyczny:
   $W_{0.05} = [\chi^2_{0.95,1}, +\infty) = [3.84, +\infty) \implies t \in W_{0.05}$.
2. Wyznaczamy p-wartość: $P(T > t|H_0) = 1 - F_{\chi^2_1}(t) = 0.0001$.

Wniosek: Na poziomie istotności 0.05 mamy podstawy do odrzucenia hipotezy zerowej, stąd uznajemy, że istnieje związek między dochodami a preferencjami politycznymi.

Dokładny test Fishera

- test niezależności stosowany zamiast testu χ^2, gdy liczebności w komórkach w tabeli są mniejsze niż 5, a całkowita liczba obserwacji jest nie większa niż 20;
- stosowany dla danych dostępnych w formie tablicy 2×2;

Tablica kontyngencji:

	B_1	B_2
A_1	n_{11}	n_{12}
A_2	n_{21}	n_{22}

Dokładny test Fishera

W ramach testu Fishera obliczane jest prawdopodobieństwo otrzymania danego rozkładu z tablicy. Rozpatrywane są wszelkie możliwe kombinacje liczebności komórek w oparciu o liczebności brzegowe zgodnie ze wzorem:

$$p = \frac{\binom{n_{11}+n_{12}}{n_{11}}\binom{n_{21}+n_{22}}{n_{21}}}{\binom{n}{n_{11}+n_{12}}}$$

Dokładny poziom istotności p jest sumą tych prawdopodobieństw, które są mniejsze lub równe badanemu prawdopodobieństwu.

Test McNemara służy do weryfikacji hipotezy o zgodności pomiędzy wynikami dwukrotnych pomiarów cechy X.

Tabela kontyngencji o wymiarach 2×2:

		Pomiar II	
		Kategoria 1	Kategoria 2
Pomiar I	Kategoria 1	O_{11}	O_{12}
	Kategoria 2	O_{21}	O_{22}

Test McNemara

Weryfikować będziemy hipotezę

$$H_0 : O_{12} = O_{21}$$

wobec hipotezy alternatywnej

$$H_1 : O_{12} \neq O_{21},$$

gdzie O_{12} i O_{21} są licznościami obserwowanymi występującymi poza główną przekątną macierzy kontyngencji 2×2, czyli licznościami mówiącymi o braku zgodności wyników dwukrotnych pomiarów.

Statystyka testowa jest dana wzorem

$$T = \frac{(O_{12} - O_{21})^2}{O_{12} + O_{21}}.$$

Statystyka ta ma asymptotycznie (dla dużych liczności) rozkład chi-kwadrat z jednym stopniem swobody.

Analiza wariancji — ANOVA

Analiza wariancji (ang. *Analysis of Variance*) służy do testowania istotności różnic między średnimi w wielu grupach. Metoda służy do oceny czy średnie wartości cechy Y różnią się istotnie pomiędzy grupami wyznaczonymi przez zmienną X.

Cel: Wyodrębnienie w całkowitej wariancji odpowiedzi Y, składniki pochodzące od poszczególnych czynników, oraz wariancję, za którą odpowiedzialny jest błąd.

Założenia:

- **Niezależność** obserwacji.
- Rozkład obserwacji w każdej z analizowanych grup powinien być zbliżony do **rozkładu normalnego**.
- **Homogeniczność wariancji** (równość wariancji): porównywane grupy nie różnią się zmiennością.

Jednoczynnikowa analiza wariancji — przykład

Związki chemiczne stosowane w leczeniu nowotworów mogą powodować obniżenie poziomu hemoglobiny we krwi (niedokrwistość). W przypadku pewnego związku chemicznego stosowanego w leczeniu nowotworów (Lek A) podejrzewano, że przy długotrwałym stosowaniu powoduje niedokrwistość (stężenie hemoglobiny we krwi poniżej 11g/dl) w większym stopniu niż inne leki tego typu.

Do badania włączono grupę 21 osób z rozpoznaniem nowotworu. W momencie przystąpienie do badania u wszystkich pacjentów poziom hemoglobiny we krwi był prawidłowy. Po zakończonej obserwacji u pacjentów ponownie wykonano morfologię krwi.

Jednoczynnikowa analiza wariancji — przykład

Wyniki badania poziomu hemoglobiny u badanych były następujące:

Lek A	10.2	8.7	12.5	13.8	7.6	8.2	9.8
Lek B	14.3	14.1	17.0	13.2	11.6	10.9	9.3
Lek C	10.4	12.0	13.6	13.5	14.7	15.3	14.9

Czy pacjenci przyjmujący lek A po zakończeniu terapii mieli istotnie inny poziom hemoglobiny we krwi niż pacjenci leczeni innymi lekami?

Jednoczynnikowa analiza wariancji — plan zrównoważony

Każda obserwacja przedstawiona jest jako suma efektów czynników, jakie zostały uwzględnione w analizie zmienności:

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

gdzie

- i oznacza numer poziomu, $i = 1, \dots, r$ (r poziomów czynnika),
- j oznacza numer obserwacji na i-tym poziomie, $j = 1, \dots, n$ (n obserwacji na każdym poziomie),
- Y_{ij} oznacza wartość cechy u j-tego obiektu pochodzącego z i-tej grupy,
- μ_i oznacza charakterystyczną wartość dla i-tego czynnika,
- ε_{ij} oznacza błąd losowy taki, że $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma)$.

Jednoczynnikowa analiza wariancji — plan zrównoważony

Model można zapisać w innej postaci:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

gdzie

- μ oznacza średnią ogólną, obliczoną dla całej populacji,
- $\alpha_i = \mu_i - \mu$ oznacza efekt *i*-tej grupy, tj. różnica między średnią dla *i*-tej grupy i dla całej populacji. Można ten efekt traktować jako przewagę *i*-tej grupy nad przeciętną dla całej populacji.

Jednoczynnikowa analiza wariancji — plan zrównoważony

Wówczas testujemy hipotezę zerową

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

lub

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$$

wobec hipotezy alternatywnej orzekającej, że istnieje co najmniej jedna para średnich/efektów, które różnią się ze sobą.

Jednoczynnikowa analiza wariancji — plan zrównoważony

Wówczas testujemy hipotezę zerową

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

lub

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$$

wobec hipotezy alternatywnej orzekającej, że istnieje co najmniej jedna para średnich/efektów, które różnią się ze sobą.

Ważne: ANOVA nie pozwala na stwierdzenie między którymi grupami występują różnice. Aby to stwierdzić konieczne jest wykonanie porównań wielokrotnych („*post-hoc*”).

Oznaczmy

- średnią ze wszystkich obserwacji

$$\overline{\overline{Y}} = \frac{1}{nr} \sum_{i=1}^{r} \sum_{j=1}^{n} Y_{ij},$$

- średnia na i-tym poziomie, $1 \leq i \leq n$

$$\overline{Y}_i = \frac{1}{n} \sum_{j=1}^{n} Y_{ij},$$

- zmienność całkowitą w eksperymencie

$$\mathrm{SST} = \sum_{i=1}^{r} \sum_{j=1}^{n} (Y_{ij} - \overline{\overline{Y}})^2,$$

- zmienność między poziomami

$$\mathrm{SSA} = n\sum_{i=1}^{r}(\overline{Y}_i - \overline{\overline{Y}})^2,$$

- zmienność na danym poziomie

$$\mathrm{SSE} = \sum_{i=1}^{r}\sum_{j=1}(Y_{ij} - \overline{Y}_i)^2.$$

- zmienność między poziomami

$$\mathsf{SSA} = n\sum_{i=1}^{r}(\overline{Y}_i - \overline{\overline{Y}})^2,$$

- zmienność na danym poziomie

$$\mathsf{SSE} = \sum_{i=1}^{r}\sum_{j=1}(Y_{ij} - \overline{Y}_i)^2.$$

Twierdzenie. $\mathsf{SST} = \mathsf{SSA} + \mathsf{SSE}$.

Jednoczynnikowa analiza wariancji — statystyka testowa

Statystyka testowa jest postaci:

$$F = \frac{\frac{\mathsf{SSA}}{r-1}}{\frac{SSE}{r(n-1)}}.$$

Przy założeniu prawdziwości hipotezy zerowej statystyka F ma rozkład F-Snedecora ze stopniami swobody $r-1$ oraz $r(n-1)$.

Duże wartości statystyki świadczą o nieprawdziwości H_0, stąd obszar krytyczny testu F jest postaci

$$[F_{1-\alpha}^{(r-1)(r(n-1))}, +\infty).$$

Zmienna losowa X ma rozkład F-Snedecora z parametrami $m, n \in \mathbb{N}_+$, jeżeli jej gęstość f jest postaci

$$f(x) = \begin{cases} \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{n}{2}-1} \left(x + \frac{m}{n}\right)^{-\frac{n+m}{2}} & \text{dla} \quad x \geq 0 \\ 0 & \text{dla} \quad x < 0, \end{cases}$$

Wartość oczekiwana i wariancja dane są wzorami

$$\mathbb{E}X = \frac{m}{m-2}, \quad \mathrm{Var}(X) = \frac{2m^2(n+m-2)}{n(m-2)(m-4)}.$$

Tabela jednoczynnikowej ANOVY

Źródło zmienności	**Suma kwadratów odchyleń**	**Stopnie swobody**	**Średni kwadrat odchyleń**	**Statystyka testowa**
Zmienność międzygrupowa (wpływ czynnika)	SSA	$r-1$	$\mathsf{MSA} = \frac{\mathsf{SSA}}{r-1}$	F
Zmienność wewnątrzgrupowa (błędy losowe)	SSE	$r(n-1)$	$\mathsf{MSE} = \frac{\mathsf{SSe}}{r(n-1)}$	F
Ogółem	SST	$rn-1$	—	—

Jednoczynnikowa ANOVA — przykład

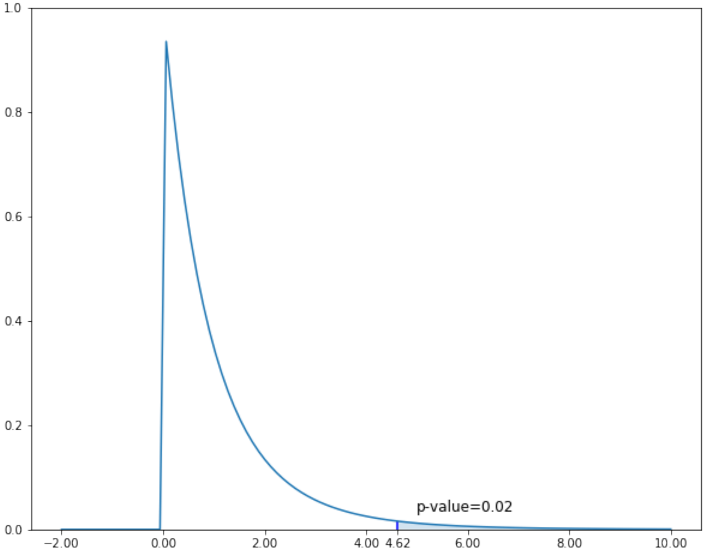
Źródło zmienności	Suma kwadratów odchyleń	Stopnie swobody	Średni kwadrat odchyleń	Statystyka testowa
Zmienność międzygrupowa (wpływ czynnika)	SSA $= 45.58$	2	MSA $= 22.79$	$F = 4.62$
Zmienność wewnątrzgrupowa (błędy losowe)	SSE $= 88.83$	18	MSE $= 4.93$	$F = 4.62$
Ogółem	SST $= 134,41$	20	—	—

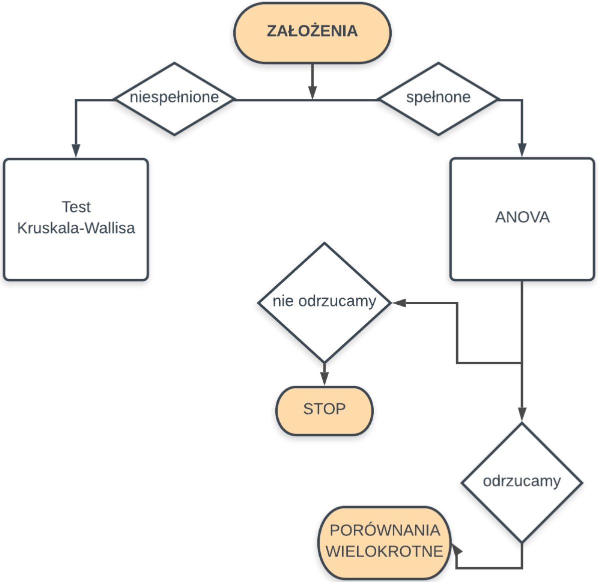
Jednoczynnikowa ANOVA — przykład

Źródło zmienności	**Suma kwadratów odchyleń**	**Stopnie swobody**	**Średni kwadrat odchyleń**	**Statystyka testowa**
Zmienność międzygrupowa (wpływ czynnika)	$\mathsf{SSA} = 45.58$	2	$\mathsf{MSA} = 22.79$	$F = 4.62$
Zmienność wewnątrzgrupowa (błędy losowe)	$\mathsf{SSE} = 88.83$	18	$\mathsf{MSE} = 4.93$	$F = 4.62$
Ogółem	$\mathsf{SST} = 134,41$	20	—	—

Wyznaczamy $p - \text{value} = P(F > 4.62) = 1 - F(4.62) = 0.02.$

Jednoczynnikowa ANOVA — przykład




ZAŁOŻENIA
niespełnione
spełnone
Test
Kruskala-Wallisa
ANOVA
nie odrzucamy
STOP
odrzucamy
PORÓWNANIA
WIELOKROTNE

Jeżeli hipoteza zerowa zostanie odrzucona, wtedy należy wykonać badania różnic średnich pomiędzy kolejnymi grupami. Służą do tego testy **post-hoc**.

Testujemy

$$H_{0,ik} : \mu_i = \mu_k,$$
$$H_{1,ik} : \mu_i \neq \mu_k$$

dla wszystkich par $i, k = 1, \ldots, r, i < k$.

Hipotezę zerową odrzucamy, gdy $|\overline{Y}_i - \overline{Y}_k|$ jest „za duża”.

Odrzucamy hipotezę zerową $H_{0,ik}$, gdy

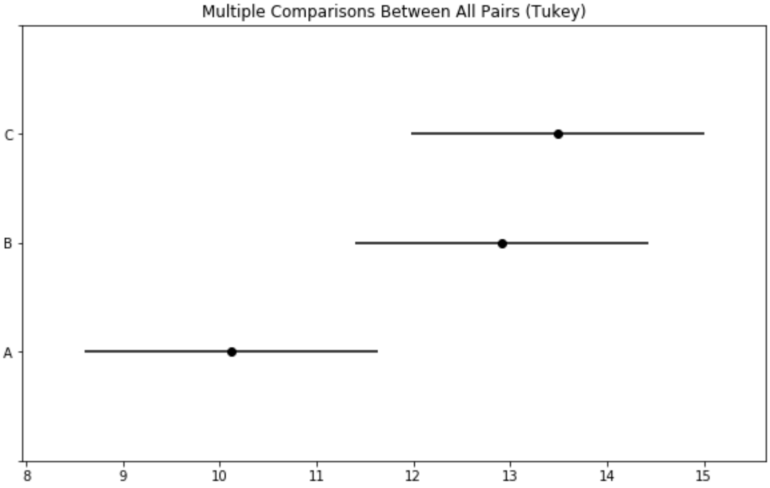
$$|\overline{Y}_i - \overline{Y}_k| \geq q_{1-\alpha}^{(r,r(n-1))}\sqrt{\frac{\mathrm{MSE}}{n}},$$

gdzie $q_{1-\alpha}^{(r,r(n-1)}$ jest kwantylem studentyzowanego rozkładu rozstępu.

Jednoczynnikowa ANOVA — przykład

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
===================================================
group1 group2 meandiff p-adj   lower  upper  reject
---------------------------------------------------
     A      B      2.8 0.0729 -0.2293 5.8293  False
     A      C   3.3714 0.0279  0.3421 6.4008   True
     B      C   0.5714 0.8772 -2.4579 3.6008  False
---------------------------------------------------
```

Jednoczynnikowa ANOVA — przykład



Pacjentów z reumatoidalnym zapaleniem stawów poproszono o ogólną ocenę stanu zdrowia w skali od 0 do 100, gdzie 0 oznacza bardzo dobre samopoczucie, a 100 bardzo złe samopoczucie. Do badania włączono 30 pacjentów, którzy aktywność choroby oceniali w granicach 70-80.

W sposób losowy wybrano po 10 pacjentów, którym podano Lek I, Lek II oraz placebo. Z każdej 10 wybrano (na drodze losowania) 5 chorych, u których równocześnie prowadzono fizjoterapię.

Dwuczynnikowa ANOVA — przykład

Po miesiącu terapii chorych poproszono o ponowne dokonanie oceny samopoczucia. Otrzymano następujące wyniki:

	Placebo	**Lek I**	**Lek II**
Prowadzono fizjoterapię	60, 54, 88, 76, 72	48, 73, 39, 35, 51	43, 67, 53, 48, 49
Nie prowadzono fizjoterapii	90, 87, 67, 55, 82	56, 76, 62, 44, 52	57, 75, 78, 64, 82

Czy na ocenę samopoczucia pacjentów miały wpływ:

- *rodzaj przyjmowanego leku,*
- *fizjoterapia,*
- *współdziałanie fizjoterapii i przyjmowanego leku?*

Dwuczynnikowa analiza wariancji

Rozważmy model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

gdzie

- μ oznacza średnią ogólną, obliczona dla całej populacji,
- α_i oznacza stały efekt *i*-tej grupy dla czynnika I,
- β_j oznacza stały efekt *j*-tej grupy dla czynnika II,
- γ_{ij} oznacza efekt interakcji pomiędzy czynnikami I i II,
- ε_{ijk} oznacza błąd losowy taki, że $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma)$.

Dwuczynnikowa analiza wariancji

Testujemy hipotezy

$$H_{0,1}: \alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$$
$$H_{1,1}: \neg H_{0,1},$$

$$H_{0,2}: \beta_1 = \beta_2 = \ldots = \beta_s = 0$$
$$H_{1,2}: \neg H_{0,2},$$

$$H_{0,3}: \gamma_{11} = \ldots = \gamma_{rs} = 0$$
$$H_{1,3}: \neg H_{0,3}$$

Tablica dwuczynnikowej ANOVY

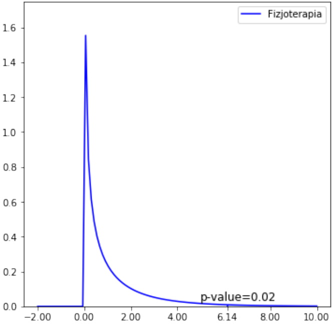
Źródło zmienności	Suma kwadratów ochyleń	Stopnie swobody	Średni kwadrat ochyleń	Statystyki testowe
Czynnik A	SSA	$r-1$	$\mathrm{MSA} = \frac{\mathrm{SSA}}{r-1}$	$F_1 = \frac{\mathrm{MSA}}{\mathrm{MSE}}$
Czynnik B	SSB	$s-1$	$\mathrm{MSB} = \frac{\mathrm{SSB}}{s-1}$	$F_2 = \frac{\mathrm{MSB}}{\mathrm{MSE}}$
Interakcje	SSAB	$(r-1)(s-1)$	$\mathrm{MSAB} = \frac{\mathrm{SSAB}}{(r-1)(s-1)}$	$F_3 = \frac{\mathrm{MSAB}}{\mathrm{MSE}}$
Błąd losowy	SSE	$rs(n-1)$	$\mathrm{MSE} = \frac{\mathrm{SSE}}{rs(n-1)}$	
Ogółem	SST	$rsn-1$		

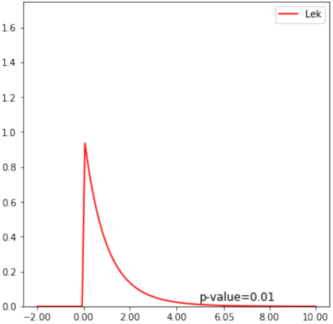
Dwuczynnikowa ANOVA — przykład

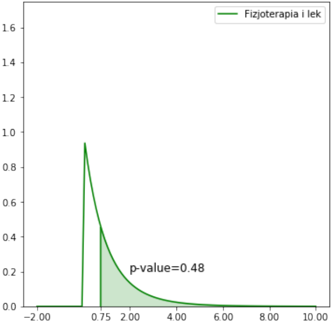
Źródło zmienności	**Suma kwadratów ochyleń**	**Stopnie swobody**	**Średni kwadrat ochyleń**	**Statystyki testowe**
Fizjoterapia	$\mathsf{SSA} = 974.70$	1	$\mathsf{MSA} = 974.70$	$F_1 = 6.14$
Lek	$\mathsf{SSB} = 1921.67$	2	$\mathsf{MSB} = 960.83$	$F_2 = 6.05$
Fizjoterapia i lek	$\mathsf{SSAB} = 236.60$	2	$\mathsf{MSAB} = 118.30$	$F_3 = 0.75$
Błąd losowy	$\mathsf{SSE} = 3810.40$	24	$\mathsf{MSE} = 158.77$	

Wyznaczamy: $p - \text{value}_{F_1} = 0.02$, $p - \text{value}_{F_2} = 0.01$, $p - \text{value}_{F_3} = 0.48$

Dwuczynnikowa ANOVA — przykład







Materiały na podstawie

- Grzegorzewski P., Bobecka K., Dembińska A., Pusz J., Rachunek prawdopodobieństwa i statystyka, WSISiZ, Warszawa, wyd. V - 2008.
- J. Koronacki, J. Mielniczuk, Statystyka dla studentów kierunków technicznych i przyrodniczych, Wydawnictwa Naukowo-Techniczne, Warszawa 2001.