# Porównanie wielu rozkładów normalnych

Założenia:
1. $X_i \sim N(\mu_i, \sigma_i^2), \quad i = 1, \ldots, k$
2. $X_1, \ldots, X_k$ są niezależne

Czy $\mu_1 = \cdots = \mu_k$?

Czy $\sigma_1^2 = \cdots = \sigma_k^2$?

Próby: $X_{i1}, \ldots, X_{in_i}, \quad i = 1, \ldots, k$

$$\bar{X}_i, \quad \mathrm{var}X_i, \quad s_i^2 = \frac{\mathrm{var}X_i}{n_i - 1}; \quad i = 1, \ldots, k$$

$$H_0 : \mu_1 = \cdots = \mu_k$$

Założenie $\sigma_1^2 = \cdots = \sigma_k^2$

**Test** $F$ (poziom istotności $\alpha$)

Statystyka testowa

$$F_{\text{emp}} = \frac{S_a^2}{S_e^2}$$

$$S_a^2 = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2$$

$$S_e^2 = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \quad \bar{X} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$$

$$N = \sum_{i=1}^{k} n_i$$

Jeżeli $F_{\text{emp}} > F\left(\alpha; k-1, N-k\right)$,
to hipotezę $H_0 : \mu_1 = \cdots = \mu_k$ odrzucamy.

**Wniosek praktyczny:**
przynajmniej jedna ze średnich $\mu_1, \ldots, \mu_k$ jest inna od pozostałych

## Model analizy wariancji

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

Błąd losowy $\varepsilon_{ij} \sim N(0, \sigma^2)$

## Przykłady

Plenność kilku *odmian* pewnej rośliny uprawnej
Wydajność pracowników kilku *zakładów* pracy
Zarobki kilku *grup* społecznych

**Czynnik:** odmiana, zakład, grupa
**Poziomy czynnika:** badane odmiany, badane zakłady, badane grupy

Model analizy wariancji

$$X_{ij} = \mu + a_i + \varepsilon_{ij}$$

$a_i$ — efekt $i$–tego poziomu czynnika: $\sum_{i=1}^{k} a_i = 0$

$$H_0 : a_1 = \cdots = a_k = 0, \qquad H_0 : \sum_{i=1}^{k} a_i^2 = 0$$

**Tabela analizy wariancji**

| Źródło zmienności | Stopnie swobody | Sumy kwadratów | Średnie kwadraty | $F_{\text{emp}}$ |
|---|---|---|---|---|
| Czynnik | $k-1$ | $\text{var}A$ | $S_a^2 = \frac{\text{var}A}{k-1}$ | $S_a^2/S_e^2$ |
| Błąd losowy | $N-k$ | $\text{var}E$ | $S_e^2 = \frac{varE}{N-k}$ | |
| Ogółem | $N-1$ | $\text{var}T$ | | |

$$\text{var}A = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2, \text{var}E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2,$$

$$\text{var}T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2,$$

$$\text{var}A + \text{var}E = \text{var}T$$

**Grupy jednorodne** — podzbiory średnich, które można uznać za takie same

**Procedury porównań wielokrotnych** — postępowanie statystyczne zmierzające do podzielenia zbioru średnich na grupy jednorodne

Procedury: Tukeya, Scheffégo, Bonfferroniego, Duncana, Newmana–Kuelsa i inne.

Ogólna idea procedur porównań wielokrotnych
$(n_1 = \cdots = n_k)$

$NIR$ — najmniejsza istotna różnica

Jeżeli $|\bar{X}_i - \bar{X}_j| < NIR$, to uznajemy, że $\mu_i = \mu_j$.

Jeżeli
$|\bar{X}_i - \bar{X}_j| < NIR$
$|\bar{X}_i - \bar{X}_l| < NIR$
$|\bar{X}_l - \bar{X}_j| < NIR$,
to uznajemy, że $\mu_i = \mu_j = \mu_l$.

Badając w ten sposób wszystkie pary średnich próbkowych otrzymujemy podział zbioru średnich na grupy jednorodne.

## Procedura Tukeya

Założenie: $n_1 = \cdots = n_k = n$

$$NIR = t(\alpha; k, N-k) S_e \sqrt{\frac{1}{n}}$$

$t(\alpha; k, N-k)$ — wartość krytyczna studentyzowanego rozstępu

Przypadek nierównolicznych prób
Jedna z modyfikacji procedury Tukeya

$$NIR_{ij} = t(\alpha; k, N-k) S_e \sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

**Przykład.** Przeprowadzić analizę porównawczą wyników punktowych klasówki w grupach studenckich.

---

**Populacje**
Możemy wyodrębnić dziesięć populacji indeksowanych numerami grup studenckich

**Cecha $X$**
Ilość punktów uzyskanych na klasówce

**Założenia**
cecha $X$ ma w $i$–tej populacji rozkład $N(\mu_i, \sigma_i^2)$
$(i = 1, \ldots, 10)$
$\sigma_1^2 = \cdots = \sigma_{10}^2$

**Formalizacja**
weryfikacja hpotezy $H_0 : \mu_1 = \cdots = \mu_{10}$

**Techniki statystyczna**
• Jednoczynnikowa analiza wariancji
• Porównania szczegółowe
Poziom istotności 0.05

## Obliczenia

| $i$ | $n_i$ | $\sum x_i$ | $\sum x_i^2$ |
|---|---|---|---|
| 1 | 30 | 18.230 | 11.375950 |
| 2 | 30 | 16.672 | 9.596790 |
| 3 | 30 | 14.292 | 7.087458 |
| 4 | 30 | 18.879 | 12.069655 |
| 5 | 30 | 18.200 | 11.355982 |
| 6 | 30 | 19.568 | 13.172884 |
| 7 | 30 | 16.522 | 9.420960 |
| 8 | 30 | 19.134 | 12.514874 |
| 9 | 30 | 18.548 | 11.945964 |
| 10 | 30 | 16.521 | 9.304785 |
| | 300 | 176.566 | 107.845302 |

| $i$ | $n_i$ | $\bar{x}_i$ | $n_i(\bar{x}_i - \bar{x})^2$ | $\mathrm{var}x_i$ |
|---|---|---|---|---|
| 1 | 30 | 0.607667 | 0.010960 | 0.298187 |
| 2 | 30 | 0.555733 | 0.032315 | 0.331604 |
| 3 | 30 | 0.476400 | 0.377351 | 0.278749 |
| 4 | 30 | 0.629300 | 0.049809 | 0.189100 |
| 5 | 30 | 0.606667 | 0.009843 | 0.314649 |
| 6 | 30 | 0.652267 | 0.121782 | 0.409330 |
| 7 | 30 | 0.550733 | 0.042911 | 0.321744 |
| 8 | 30 | 0.637800 | 0.072757 | 0.311209 |
| 9 | 30 | 0.618267 | 0.026486 | 0.478354 |
| 10 | 30 | 0.550700 | 0.042986 | 0.206670 |
| | $N$=300 | $\bar{x}$=0.588553 | var$A$=0.787199 | var$E$=3.139595 |

$$\mathrm{var}T = 107.845302 - 176.566^2/300 = 3.926794$$

## Tabela analizy wariancji

| Źródło zmienności | Stopnie swobody | Sumy kwadratów | Średnie kwadraty | $F_{\text{emp}}$ |
|---|---|---|---|---|
| Grupa | 9 | 0.787199 | 0.087467 | 8.079 |
| Błąd losowy | 290 | 3.139595 | 0.010826 | |
| Ogółem | 299 | 3.926794 | | |

Wartość krytyczna

$$F\left(0.05; 9, 290\right) = 1.912$$

**Odpowiedź:**
hipotezę $H_0 : \mu_1 = \cdots = \mu_{10}$ odrzucamy

**Wniosek:**
przynajmniej jedna grupa uzyskała inną średnią liczbę punktów niż pozostałe

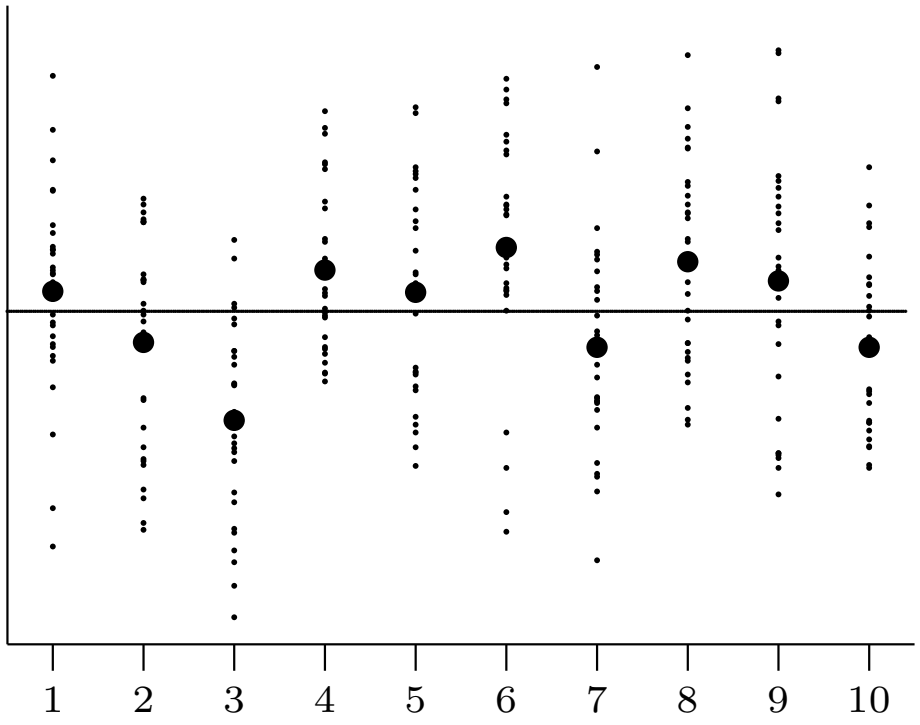
## Wyznaczenie grup jednorodnych

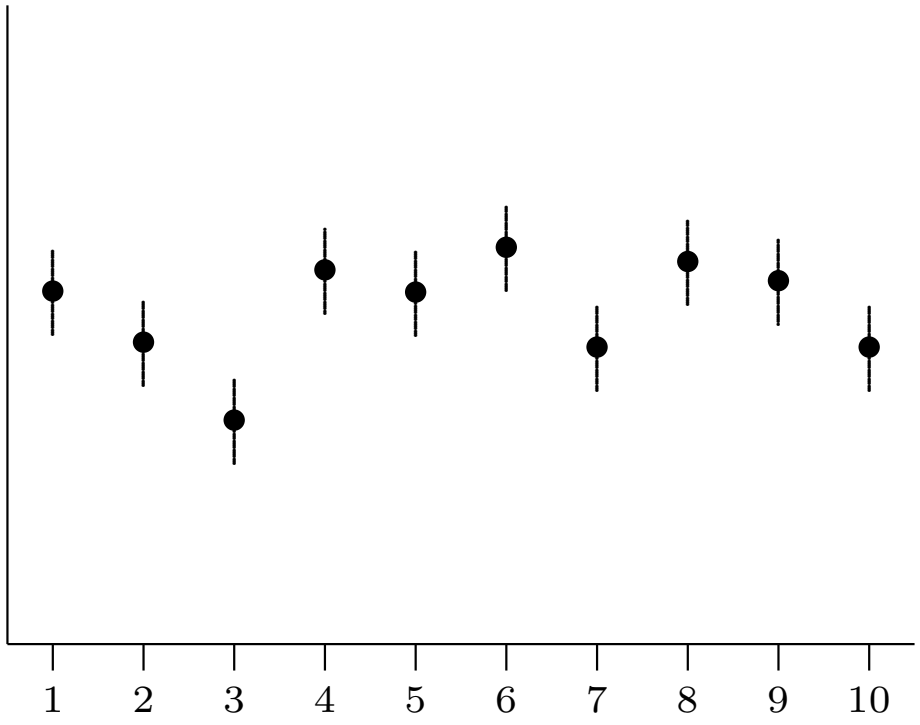
Procedura Tukeya ($\alpha = 0.05$)

Wartość krytyczna: $t(0.05; 10, 290) = 4.474$

$$NIR = 4.474 \cdot \sqrt{0.010826} \cdot \sqrt{\frac{1}{30}} = 0.084990$$

| $i$ | $\bar{x}_i$ | | | | |
|---|---|---|---|---|---|
| 3 | 0.476400 | * | | | |
| 10 | 0.550700 | * | * | | |
| 7 | 0.550733 | * | * | | |
| 2 | 0.555733 | * | * | * | |
| 5 | 0.606667 | | * | * | * |
| 1 | 0.607667 | | * | * | * |
| 9 | 0.618267 | | * | * | * |
| 4 | 0.629300 | | * | * | * |
| 8 | 0.637800 | | | * | * |
| 6 | 0.652267 | | | | * |


1
2
3
4
5
6
7
8
9
10


1
2
3
4
5
6
7
8
9
10

**Porównanie wariancji**
Cecha $X_i$ ma rozkład normalny $N(\mu_i, \sigma_i^2)$
Średnie $\mu_i$ oraz wariancje $\sigma_i^2$ są nieznane

$$H_0 : \sigma_1^2 = \cdots = \sigma_k^2$$

**Test Bartletta** (poziom istotności $\alpha$)
Statystyka testowa

$$M = (N-k) \ln \left( \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1) S_i^2 \right) - \sum_{i=1}^{k} (n_i - 1) \ln S_i^2$$

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

Jeżeli $M > m(\alpha)$, to $H_0 : \sigma_1^2 = \cdots = \sigma_k^2$ odrzucamy.

$$m(\alpha) =$$
$$\frac{1}{c_1 - c}[(c_1 - c_3)m_1(\alpha; k, c_1) + (c_3 - c)m_2(\alpha; k, c_1)]$$

$$c_1 = \sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{N - k}$$

$$c_3 = \sum_{i=1}^{k} \frac{1}{(n_i - 1)^3} - \frac{1}{(N - k)^3} \ ,$$

$$c = c_1^3/k^2 \ ,$$

$m_1(\alpha; k, c_1), \quad m_2(\alpha; k, c_1)$ są stablicowane
Jeżeli wszystkie $n_i > 4$, to statystyka testowa

$$\frac{M}{1 + c_1/(3(k-1))}$$

ma w przybliżeniu rozkład chi–kwadrat z $k-1$ stopniami swobody.
Jeżeli $c_1 = 0$, to

$$m_1(\alpha; k, c_1) = m_2(\alpha; k, c_1) = \chi^2(\alpha; k-1)$$

Przypadek $n_1 = \cdots = n_k = n$

**Test Cochrana** (poziom istotności $\alpha$)
Statystyka testowa

$$G = \frac{S^2_{\max}}{S^2_1 + \cdots + S^2_k}$$

$$S^2_{\max} = \max\{S^2_1, \ldots, S^2_k\}$$

Jeżeli $G > g(\alpha; k, n)$,
to $H_0 : \sigma^2_1 = \cdots = \sigma^2_k$ odrzucamy

Wartości krytyczne $g(\alpha; k, n)$ są podane w tablicach

Przypadek $n_1 = \cdots = n_k = n$

**Test Hartleya** (poziom istotności $\alpha$)
Statystyka testowa

$$F_{\max} = \frac{S^2_{\max}}{S^2_{\min}}$$

$$S^2_{\min} = \min\{S^2_1, \ldots, S^2_k\}$$

Jeżeli $F_{\max} > f_{\max}(\alpha; k, n)$,
to $H_0 : \sigma^2_1 = \cdots = \sigma^2_k$ odrzucamy

Wartości krytyczne $f_{\max}(\alpha; k, n)$ są podane w tablicach

**Przykład.** W celu porównania zróżnicowania cen targowiskowych na jaja w czterech województwach w Polsce z każdego województwa wylosowano pewne ilości targowisk i zanotowano przeciętne ceny jaj na tych targowiskach. Po odpowiednich przeliczeniach uzyskano nastepujące wyniki

| Województwo | Liczba targowisk $n_i$ | Wariancja $s_i^2$ |
|---|---|---|
| 1 | 8 | 900 |
| 2 | 6 | 400 |
| 3 | 5 | 400 |
| 4 | 7 | 1600 |

Czy można na tej podstawie uznać, że zróżnicowanie cen w badanych województwach jest takie same?

---

**Populacje**
Są cztery populacje: targowiska w badanych województwach

**Cecha** $X$
przeciętna cena jaj na targowisku

**Założenie**
cecha w $i$–tej populacji ma rozkład $N(\mu_i, \sigma_i^2)$
$(i = 1, 2, 3, 4)$

**Formalizacja**

Miernikiem zróżnicowania cechy jest jej wariancja. Zatem problem analizy porównawczej zróżnicowania cen można zapisać jako zagadnienie weryfikacji hipotezy $H_0 : \sigma_1^2 = \cdots = \sigma_4^2$

**Technika statystyczna**

Test Bartletta (poziom istoności $\alpha = 0.05$)

**Obliczenia**

| | $n_i$ | $(n_i-1)s_i^2$ | $(n_i-1)\ln s_i^2$ | $1/(n_i-1)$ | $1/(n_i-1)^3$ |
|---|---|---|---|---|---|
| 1 | 8 | 6300 | 47.6168 | 0.1429 | 0.0029 |
| 2 | 6 | 2000 | 29.9573 | 0.2000 | 0.0080 |
| 3 | 5 | 1600 | 23.9659 | 0.2500 | 0.0156 |
| 4 | 7 | 9600 | 44.2666 | 0.1667 | 0.0046 |
| Razem | 26 | 19500 | 145.8065 | 0.7595 | 0.0312 |

$$M = (26-4)\ln\left(\frac{19500}{26-4}\right) - 145.8065 = 3.5103$$

$$c_1 = 0.7595 - \frac{1}{(26-4)} = 0.7141$$

$$c_3 = 0.0312 - \frac{1}{(26-4)^3} = 0.0311$$

$$c = \frac{0.7141^3}{4^2} = 0.0228$$

Wartość krytyczna

$$m_1(0.05; 4, 0.7141) = 8.4630$$

$$m_2(0.05; 4, 0.7141) = 8.0972$$

$$m(0.05) = \frac{(0.7141-0.0311)8.4630+(0.0311-0.0228)8.0972}{0.7141-0.0228} = 8.4586$$

**Odpowiedź:** nie ma podstaw do odrzucenia weryfikowanej hipotezy

**Wniosek:** zróżnicowanie cen targowiskowych w badanych województwach można uznać za takie same.