

# ---== STATYSTYKA ==---

## ---== Temat 3 ==---

**Populacja:** zbiór/rzecz którą chcemy przebadac np.: ludzie mieszkający w PL lub populacja gwoździ w fabryce.

**Zmienna/cecha:** określona rzecz/charakterystyka którą badamy w populacji. Każda cecha ma swój rozkład prawdopodobieństwa. Bardzo często, nie jesteśmy w stanie zbadać całej populacji, dlatego też pracujemy na tak zwanej **próbie**: czyli podzbiorze populacji.

Dane - wektor:  $X = (y_1, \dots, y_n)^T$  [ gdzie X: jest to cecha/zmienna || a Y: jest to pojedyncza obserwacja ]

Metoda analizy danych jest zależna od typu obserwacji (obserwacja jakościowa czy ilościowa?)

Celem statystyki opisowej jest przedstawienie rozkładu empirycznego badanej zmiennej X na danej populacji/próbie za pomocą wykresu, tabeli lub liczb (opisujemy w ten sposób obserwacje - dlatego „statystyka opisowa”)

Metody graficznego opisu rozkładu empirycznego:

- wykres słupkowy (jakościowa lub dyskretna zmienna ilościowa)
- wykres kołowy (jakościowa lub dyskretna zmienna ilościowa)
- histogram (ciągła zmienna ilościowa)
- wykres pudełkowy lub ramka-wąsy lub ramkowy (zmienna ilościowa)

**STATYSTYKI OPISOWE (rozkładu empirycznego):**

**MIARY TENDENCJI CENTRALNEJ** - lokalizują środek zbioru danych - wartości średnie:

- **$\bar{X}$ : Średnia arytmetyczna** jest dobrą miarą o ile mamy **rozkład symetryczny**. Jeśli natomiast mamy odstępstwa od tej symetrii (obserwacje odstające) wtedy średnia może być przez nie zniekształcona.
- **Me: Mediana** (szczególny przypadek kwantylu rzędu 1/2) bazuje na próbie uporządkowanej (posortowanych wartościach od najmniejszego do największego) Połowa obserwacji jest większa od mediany i połowa

obserwacji jest mniejsza od mediany. Mediana jest bardziej odporna na pojawienie się obserwacji odstających. (estymator odporny - odporny na obserwacje odstające)

- Przy rozkładzie **symetrycznym** **średnia i mediana będą miały podobne wartości** ale przy rozkładzie asymetrycznym - będą się już bardziej od siebie różnić.
- **$Q_p$  Kwantyle / kwartyle** - można tworzyć je dla każdego rzędu. (1/4 np.: wskazuje nam podział, że 25% wartości jest mniejsze od  $K1/4$  a 75% jest większe

**MIARY DYSPEKSJI** rozkładu empirycznego (miary rozproszenia/rozrzutu - określamy je po określeniu środka zbioru danych):

- Tutaj patrzemy jak dane są skoncentrowane/zgrupowane czyli: czy dane obserwacji raczej są blisko wartości średniej czy raczej są bardziej rozrzucone. Im większa jest ta miara tym większy jest rozrzut.
- **S: Odchylenie standardowe** (mierzy rozrzut wartości obserwacji od średniej) Odchylenie standardowe (s) jest o tyle dobre, bo praktycznie bardzo łatwo jest go zinterpretować, ponieważ ma jednostki zwykłe (jak coś mierzymy w cm. To wynik również będzie w cm.) W przeciwieństwie do wariancji  **$s^2$**  - która jest trudniejsza w interpretacji ale w teorii jest fajniejsza.
- **$S^2$**  łatwiej się posługiwać wariancją ( **$s^2$** ), a interpretować łatwiej odchylenie standardowe.
- **V: Współczynnik zmienności** - podajemy w procentach. Jest to odchylenie standardowe podzielone przez średnią. Jego zaletą jest brak jednostki i pozwala nam na porównywanie zmienności tych samych cech tylko mierzonych różnymi sposobami. Im większy - tym większa zmienność!

## MIARY ASYMETRII ROZKŁADU:

- **A: Współczynnik asymetrii** pozwala na liczbowe scharakteryzowanie odchylenia standardowego.  $0 = A \rightarrow$  symetria |  $A > 0 \rightarrow$  prawostronna asymetria |  $A < 0$  lewostronna asymetria.

Pole które występuje pod krzywą wskazuje nam na prawdopodobieństwo otrzymania konkretnych wartości. Np.: więc w przypadku prawostronnej asymetrii, bardziej prawdopodobne jest wystąpienie wartości mniejszych w naszej czesze. Im większa wartość tym to pole znacznie się zmniejsza - a więc również zmniejsza się szansa na wystąpienie większych wartości w danej czesze.

Średnia w przypadku prawostronnej asymetrii jest większa od mediany. A W przypadku lewostronnej - mniejsza. (Mediana dlatego jest lepsza - bo średnia jest zawyżana/zaniżana: przez odstające wartości)

- **K: kurtoza (Współczynnik koncentracji rozkładu)** bazuje na 4potęgach odchylenia obserwacji. Ona mówi nam o występowaniu obserwacji skrajnych/odstających - a dokładniej o grubości ogonów obserwacji odstających.

Rozkład normalny ma cienkie ogony  $\rightarrow$  kurtoza = 0

Rozkład jednostajny - ma zanik ogonów (wynik na minusie - bo ogony są węższe niż w rozkładzie normalnym i jest mniejsze ryzyko wystąpienia obserwacji odstających niż w rozkładzie normalnym)

Rozkład Cauchy'ego ma kurtoze około 200! Ma bardzo grube ogony - mają one spore pola. Dużo wartości odstających.

Wiele testów bazuje na kurtozie.  $|K| < 2$  lub  $|K| < 3$  jest to test który mówi nam że jest to taka bezpieczna wartość do stosowania testów parametrycznych które zakładają normalność.

--- R  $\rightarrow$  interpretacja odchylenia standardowego , wariancji czy współczynnika zmienności jest zależna od kontekstu! Przy zmiennych jakościowych prawie zawsze nie ma sensu liczenie średnich , median, odchyłeń itp.

Przy zmiennej ilościowej ciągłej musimy pogrupować (stworzyć przedziały/klasy) nasze obserwacje aby móc stworzyć szereg rozdzielczy.

**Wykres ramkowy/pudełkowy/ramka-wąsy** -> funkcja boxplot - raczej dla danych ilościowych ciągłych! (ewentualnie dla dyskretnych ale gdzie jest dużo możliwych wartości!)

- Pogrubiona czarna linia: (znajduje się w środku ramki/pudełka) - to jest MEDIANA
- Pudełko: ma określoną wysokość. Górna krawędź to 3Kwartył a dolna to 1Kwartył. Wysokość pudełka to ich różnica)
- Powyżej górnej krawędzi - znajduje się górny 3kwartył (czyli powyżej znajduje się 25% większych wartości a poniżej 75 % mniejszych wartości)
- A poniżej dolnej krawędzi znajduje się dolny 1Kwartył (czyli poniżej znajduje się 25% mniejszych wartości a powyżej 75% większych.) Ta wysokość to rozstęp między kwartyłowy - im wysokość jest większa tym jest większy rozrzut danych. Im prostokąt jest cieńszy tym jest mniejsza zmienność.
- Wąsy - czyli odcinki na górze i na dole które są łączone z pudełkiem linią przerywaną. Jest to odpowiednio obserwacja największa i najmniejsza. Im dłuższe są te wąsy, tym większy jest rozrzut (szerokość)
- Asymetrie wskazują takie rzeczy jak. Mediana bliżej góry a wąs górny krótszy od dolnego - asymetria lewostronna
- A jeśli: Mediana bliżej dołu i wąs na dole jest krótszy od tego na górze to mamy asymetrię prawostronną.
- Jeśli wąsy są podobnej długości i mediana jest na środku pudełka to mamy rozkład symetryczny!
- Wysokość wąsów mówi nam o ogonie w rozkładzie im dłuższa linia przerywana tym więcej wartości/obserwacji zanotowaliśmy w tym przedziale
- Kółkami oznaczamy wartości odstające.

## ---== Temat 4 ==---

**Model statystyczny** - nie jest idealny, jest to uproszczenie rzeczywistości. Na takich modelach można bazować teoretycznie i działać praktycznie bazując na konkretnym modelu.

- Konkretnie liczby, wartości obserwacji potraktujemy jako realizacje pewnego wektora losowego - to będzie nasza próba losowa.

- **Próba prosta** - zakładamy że to są niezależne zmienne losowe - i że te zmienne losowe mają ten sam rozkład.
- **P** - jest to rozkład który wybierzemy (mówimy że nasze dane pochodzą z próby o rozkładzie P (tym wybranym))
- **Modele parametryczne** - my narzucamy konkretny rozkład prawdopodobieństwa na to P.
- **Modele nieparametryczne** - nie podamy rozkładu tylko ogólną charakterystykę - czy jest to zmienna dyskretna czy ciągła.

## ROZKŁADY DYSKRETNE:

- Rozkład dwumianowy (zero - jedynkowy)
- Rozkład Poissona

## ROZKŁADY CIĄGŁE:

- Rozkład jednostajny
  - Rozkład normalny
  - Rozkład wykładniczy
  - Rozkład Rayleigha
1. W tym temacie rysujemy wykresy na realnych danych na podstawie prawdopodobieństwa!
  2. Gdy mamy rozkład **dyskretny** to patrzymy, czy wiemy jaka jest maksymalna wartość czy jej nie znamy.
  3. Gdy mamy rozkład **ciągły** to rysujemy histogram (prawdopodobieństwo) i rysujemy funkcji gęstości - czerwoną krzywą i na jej podstawie dobieramy model.
  4. Patrzymy jak się zachowuje cecha teoretycznie, jakie może mieć wartości, patrzymy jakie wartości i zachowanie jest w danych (faktycznych obserwacjach) - czy się powtarzają czy nie, patrzymy na wykresy.

---

Zagadnieniem szukania tych nieznanymi parametrów - szukaniem ich oszacowań, zajmuje się **ESTYMACJA PUNKTOWA** (estymacja czyli szacowanie)

**Estymatory** - wzory/sposoby na oszacowanie parametrów.

Estymator jest statystyką - jest niezależny od TETY - statystyka nie zależy od parametru! Bo przecież ona go nam szacuje!

$g(\theta)$  - zbiór możliwych wartości

Metoda: **estymator największej wiarygodności (ENW)** - metoda automatyczna wyprowadzania estymatorów - on znajduje nam maximum dla funkcji gęstości.

Metoda: **estymator nieobciążony (EN)** - przez średnią arytmetyczną - daje nam średnią wartość estymatora, a wariancja określa rozrzut.

Metoda: **estymatora nieobciążone o minimalnej wariancji (ENMW)** - żeby miało jak najmniejszy rozrzut, żeby był jak najbardziej precyzyjny! Wariancja powinna mieć jak najmniejszą wartość - aby mieć jak największą precyzję. Szukamy w rodzinie estymatorów ten który ma najmniejszą wariancję.

---

Jak już wyliczymy prawdopodobieństwo teoretyczne to warto zobaczyć jego sumę.

Sum(probs) - im wartość bliżej jedynki tym lepiej dostosowany model.

---

**Przedziały ufności** - wykorzystywana np.: do kontroli jakości / spełniania norm - określa się pewien zakres wartości jaki może być np.: dla długości gwoździ, że są jeszcze w normie.

Wyniki głównie znajdują się w wyznaczonym przedziale limitów ufności

**Alfa = 0.05**

Mamy przedział np.: pomiędzy 5 a 10 i w tym przedziale spodziewamy się że znajdzie się wartość naszego parametru np. 7 i to powinno zajść z dużym prawdopodobieństwem.  $\geq 1 - \alpha$  (0.95)

W tych przedziałach często znajdują się estymatory

Przedział ufności nie może zależeć od parametru którego estymuje.

## ---== Temat 5 ==---

- Stawiamy hipotezy i następnie sprawdzamy je na podstawie próby/populacji za pomocą testów statystycznych. Najpierw jest postawione jakieś przepuszczenie a później je weryfikujemy.
- Testowana Hipoteza jest nazywana hipotezą zerową i jest oznaczana jako  $H_0$  i zestawiamy ją z inną hipotezą  $H_1$  nazywaną hipotezą alternatywną (wszystkie inne możliwości/wyniki) Przyjmujemy  $H_1$  gdy odrzucimy  $H_0$
- Wyróżniamy hipotezy dotyczące rozkładu i dotyczące konkretnych parametrów np.: długość drogi hamowania, wariancji
- Testy statystyczne bazują na obserwacjach  $f: X \rightarrow \{0,1\}$  i dają nam 2 wartości 0 lub 1 (0 oznacza, że opowiadamy się za hipotezą zerową)
- **Poziom istotności alfa = 0.05**
- Pr. Popełnienia błędu pierwszego rodzaju jest  $\leq \alpha$  (znaczy to, że my błąd pierwszego rodzaju trzymamy w ryzach, wiemy że on może nastąpić ale w maksymalnie 5% (0.05) przypadków się pomylimy (kontrolujemy poziom błędu!) Wtedy Pr. Błędu drugiego rodzaju - MINIMALIZUJEMY! (prawo wagi szalkowej - ciężary itp.) Nie da się minimalizować błędów jednocześnie!
- Moc testu maksymalnie może przyjąć wartość 1. (w 100% przypadków nie popełnimy błędu drugiego rodzaju)
- Jeżeli P wartość jest  $\leq$  poziomowi istotności alfa to  $H_0$  odrzucamy
- Jeżeli P wartość jest  $>$  poziomowi istotności alfa to nie mamy podstaw do odrzucania  $H_0$ .

### PROCEDURA TESTOWA:

- 1) Dane  $X=(x_1, \dots, x_n)$
- 2) Obierz hipotezy
- 3) Wybierz test statystyczny i ustal poziom istotności
- 4) Oblicz p wartość
- 5) Porównaj p wartość z poziomem istotności
- 6) Podejmij decyzję

Kiedy p-wartość jest bardzo blisko wartości alfa - to nie powinniśmy być w 100% pewni że test dał nam prawidłową odpowiedź - powinniśmy nasze dane poprawić - np.: zebrać więcej obserwacji.



## RODZAJE TESTÓW:

### 1) Test normalności Shapiro-Wilka:

test pomocniczy - mówi nam czy dane pochodzą z rozkładu normalnego

$H_0$ : Gdy rozkład jest rozkładem normalnym

$H_1$ : gdy jest innym rozkładem

```
shapiro.test(x) #patrzemy na p-value względem alfy
```

### 2) Test F-Snedecora:

test pomocniczy - dwie próby niezależne. Określa nam czy dwie próby mają takie same wariancje - czy różnice między nimi są istotne.

$H_0$ : gdy wariancje są takie same

$H_1$ : gdy są różne.

```
var(x1) #porównujemy wartości 2 wariancji: < czy >  
var(x2)  
var.test(x1, x2, alternative = "less") # lub "greater" i patrzemy na p-v
```

### 3) Testy t-Studenta:

dotyczą hipotez odnośnie parametru MU (badamy średnią obserwacji) w rozkładzie normalnym. Musimy założyć normalność - dla każdej próby robimy test Shapiro-Wilka. Dla  $H_1$  najlepiej przyjąć że jest „<” lub „>” dlatego że wynik tego daje nam więcej informacji - w którym kierunku hipoteza się nie zgadza. Test t-studenta ma wtedy większą moc - tzn. błąd 2 rodzaju jest mniejszy. Wyróżniamy testy:

#### a) Dla 1 próby:

$H_0$ : MU == ustalona w zadaniu wartość Mu

$H_1$ : istotnie różne: < lub >

```
#test Shapiro-wilka  
mean(x) #porównujemy wartość z ustaloną wartością mu: < czy >  
t.test(x, mu = 250, alternative = "less") #albo "greater"
```



- b) **Dla 2 prób niezależnych:** Próby mogą mieć różną ilość obserwacji. Muszą mieć równe wariancje - wykonujemy **test F-Snedecora**. Narysuj wykres boxplot aby zobaczyć czy pewne przesłanki mają racje bytu. np. czy różnice w rozrzucie (wariacji) są istotnie różne.

$H_0: \mu_1 = \mu_2$  (średnie wartości są identyczne).

$H_1$ : istotnie różne: < lub >

```
#test Shapiro-wilka
#test F-snedecora
mean(x1)
mean(x2)
t.test(x1, x2, var.equal = TRUE, alternative = 'greater')
```

- c) **Dla 2 prób zależnych:** obserwacje muszą być równoliczne. Tutaj wariancje **mogą** być różne.

$H_0: \mu_1 = \mu_2$

$H_1$ : istotnie różne: < lub >

```
#test Shapiro-wilka
mean(x1)
mean(x2)
t.test(x1, x2, alternative = 'greater', paired = TRUE)
```

#### 4) **Test Welcha:**

dla dwóch prób niezależnych. Zakładamy normalność, dopuszczamy różne wariancje. Ale gdy mamy równość wariancji czyli  $SIG_1 = SIG_2$  to zdecydowanie lepiej wybrać test T-studenta bo on ma większą moc niż test Welcha! A gdy wariancje są różne - wtedy wybieramy test Welcha bo on da nam bardziej adekwatne wyniki.

$H_0: \mu_1 = \mu_2$

$H_1$ : istotnie różne: < lub >

```
#test Shapiro-wilka
mean(x1)
mean(x2)
t.test(x1, x2, var.equal = FALSE, alternative = 'greater')
```

POWYŻEJ -> TESTY PARAMETRYCZNE - ZAKŁADAJĄCE ROZKŁAD NORMALNY

## 5) Test Manna-Whitneya-Wilcoxona:

Tym testem możemy przebadąć każdy powyższy przypadek (badamy parametr) ale przy tym **nie** zakładamy normalności. Cecha musi być **ciągła**! Test ten bazuje na tak zwanych rangach. Obserwacje są sortowane, uporządkowane od najmniejszego do największego i nadaje im się rangi. Następnie tworzony jest wektor rang. Badanie możemy wykonać np. na średniej czy medianie.

### PRZYKŁAD - 2 niezależne próby:

```
median(x1) #może być też średnia
median(x2)
wilcox.test(x1, x2, alternative = 'greater')
```

### # Test można przeprowadzić dla pojedynczych, podwójnych zależnych i niezależnych:

```
# wilcox.test(x, y = NULL,
#             alternative = c("two.sided", "less", "greater"),
#             mu = 0, paired = FALSE, ...)
```

## 6) Test istotności dla wskaźnika struktury/proporcji:

(**prawdopodobieństwo/proporcje**)

Cecha **zero-jedynkowa, sukces-porażka**. Badamy czy coś wystąpiło czy nie. Powinno być  $\geq 100$  obserwacji do badania tym sposobem.

3 rodzaje badania wskaźnika struktury:

- d) **Dla jednego wskaźnika struktury (dyskretne)** - Liczymy prawdopodobieństwo sukcesu (jest nieznane - będziemy je estymować - średnia!) w próbie pojawią się 0 i 1, tak/nie.

$H_0: p = p_0$  (pr.)

$H_1: p (< \text{ lub } >) p_0$

```
prop.test(x = IleSukc, n = ileObse, p = ustalonePr., alternative = "less")
LUB test dwumianowy
binom.test(x = IleSukc, n = ileObse, p = ustalonePr, alternative = "less")
```

- e) **Dla dwóch wskaźników struktury (niezależne cechy)**

porównujemy prawdopodobieństwo tych dwóch struktur

$H_0: p_0 = p_1$  (pr.)

$H_1: p_0 \neq p_1$

```
prop.test(c(sukces1, 368), c(łącznieObser, 800), alternative = "greater")
```

#### f) Test McNemary: 2 wskaźniki struktury (zależne cechy)

Wyniki zero-jedynkowe

$H_0: p_1 = p_2$  (pr.)

$H_1: p_1 \neq p_2$

```
X <- matrix(c(212, 256, 144, 707), nrow = 2) #pisane kolumnami
mcnemar.test(X)
```

### 7) Testy $\chi^2$ Pearsona ( $\chi^2$ ):

Zmienna jakościowa/ilościowa która może mieć więcej niż 2 wartości. Dotyczą badania szczególnego rozkładu dyskretnego! Jeśli w tego typu testach wykorzystujemy jakiś parametr(y) to wtedy musimy obliczyć wynik za pomocą innej funkcji i zmniejszyć stopień istotności: (liczba wartości które może przyjąć zmienna --1 --ilość wykorzystanych parametrów)

#### g) test dla jednej próby dla rozkładu dyskretnego:

**sprawdzenie poprawności zaproponowanego modelu**

Wartości które mogą przyjmować zmienne jest od 1 do K.

Może być to np.: wykształcenie np.: 1.niższe, 2.średnie, 3.wyższe (nie jest to test zero-jedynkowy!!!)

$H_0: p = p_0$

$H_1: p \neq p_0$

gdzie p jest wektorem Prawdopodobieństw! Zestawiamy prawdopodobieństwo z danych z teoretycznego rozkładu.

```
lambda_est <- mean(x)
p0 <- c(dpois(0:6, lambda_est), 1 - ppois(6, lambda_est))
chisq.test(table(x), p = p0) # patrzymy na X-squared
# liczba stopni swobody = 8 - 1 - 1
1 - pchisq(2.1658, 6)
```

```
#Lub dla danych, jednej próby niezależnej
chisq.test(x=c(38,72,40),p=c(0.2,0.5,0.3))
```

#### h) test dla dwóch prób niezależnych:

```
x <- matrix(c(20, 85, 5, 39, 95, 6), nrow = 3)
chisq.test(x)
```

## 8) Testy Kołmogorowa-Smirnowa:

zakładamy że rozkład jest ciągły ale **nie** musi być normalny!

Test ten nie pozwala na estymację parametrów! One muszą być już znane.

- i) **Dla jednej próby**: Testujemy czy dystrybuanta prawdziwa  $F$  jest równa tej teoretycznej dystrybuancie  $F_0$ . Nie dobre do testowania normalności. Może ją błędnie wykazać - test Shapiro-wilka jest odnośnie tego o wiele lepszy!

$$H_0: F = F_0$$

$$H_1: F \neq F_0$$

```
set.seed(12345)
x <- runif(30) #random + nazwa rozkładu
ks.test(x, "punif") #probability + nazwa rozkładu
```

- j) **Dla 2 prób**: testujemy czy 2 rozkłady są takie same, czy pochodzą z jednej populacji. Porównujemy 2 dystrybuanty: Tutaj obie dystrybuanty są nieznane i patrzymy na odległość między nimi.

$$H_0: F = G$$

$$H_1: F \neq G$$

```
ks.test(x, y) #np.: czy 2 próby pochodzą z tej samej populacji
```

## ---== Temat 6 ==---

- Rozszerzenie testów parametrycznych analizy wariancji t-studenta z 2 niezależnych prób na 3, 4 lub więcej prób. Grupy/próby nie muszą być równoliczne - ale **MOCNO SIĘ ZALECA ŻEBY BYŁY!** Taka sytuacja jest najlepsza.
- Zmienna ciągła -  $Y_{ij}$  (zakładamy rozkład normalny ale są też testy gdy nie ma normalności)
- $Y_{ij} = \mu_i + e_{ij}$  -> gdzie  $e$  to błędy losowe - niezależne zmienne losowe o rozkładzie normalnym co mają średnią: 0 i wariancję:  $\sigma^2$   
i -> numer grupy do której należy dana obserwacja.  
j -> numer obserwacji w grupie

$H_0: \mu_1 = \mu_2 = \dots = \mu_n$  (w każdej z grup jest taka sama wartość estymatorów średniej)

$H_1: \neg H_0$  (przynajmniej w dwóch próbach będą znacząco różne średnie).

### TEST STATYSTYCZNY:

#### TEORIA:

- Test analizy wariancji - (metoda analizy wariancji) nie testuje równości wariancji, ona testuje równość średnich wartości oczekiwanych. Natomiast nazwa: „analiza wariancji” pochodzi od tego jak ten test jest konstruowany, on bierze pod uwagę wariancję, on bierze pod uwagę zmienność danych. Robi to w ten sposób, że zmienność całkowitą danych, rozbija na 2 części na część związaną z hipotezą zerową i zmienność związaną z błędem losowym. To rozbiecie pozwala nam na wykonie testu. To rozbiecie się dokonuje przez estymatory wariancji. Używamy estymatorów wariancji do konstrukcji testu dla średnich.

Musimy sprawdzić **testem Shapiro-wilka** dla reszt czy rozkłady są **normalne** oraz czy **wariancje** istotnie się nie różnią (mamy do wyboru kilka testów analizy wariancji)!

Wyróżniamy w danych 2 zmienne:

- 1) Zmienna zależna (faktyczne obserwacje -  $y$ )
- 2) Zmienna grupująca (dzieli nam dane na grupy - dobrze aby każda grupa była równoliczna)

## ALGORYTM DZIAŁANIA:

```
summary(dane) #podsumuje nasze dane | pokazuje nam średnią, medianie, min i max  
#zlicza nam ile obserwacji jest w każdej zmiennej grupującej
```

```
aggregate(zmiennaZależna, list(DOSE = zmiennaGrupująca), FUN = mean)  
#Tworzy estymatory punktowe (średnie) dla każdej z grup - zapisuje dane w tabeli
```

```
boxplot(nazwaZależnej ~ nazwaGrupującej, data = zmiennaCałeDane)  
# zmienna1 ~ zmienna2 (zmienna 1 jest modelowana przez zmienną 2)
```

**#1 TEST WŁAŚCIWY! Testy analizy wariancji:**

#  $H_0$  :  $\mu_1 == \mu_2 == \dots == \mu_x$

#  $H_1$  :  $!H_0$

#  $\Pr(>F)$  -> zawiera p-value którą musimy porównać z alfa

```
summary(aov(response ~ dose, data = x)) #test analizy wariancji
```

**#2 SPRAWDZENIE NORMALNOŚCI dla reszt! Test Shapiro-wilka**

```
shapiro.test(lm(response ~ dose, data = x)$residuals)
```

#gdy wyjdzie brak normalności - przerwij algorytm i wykonaj testy Kruskala-Wallisa

**#3 SPRAWDZENIE RÓWNOŚCI WARIACJI! Wybierz jakiś test! Jeśli jest normalność**

Np.: test Bartletta

Jeżeli w głównym teście nasza hipoteza  $H_0$  zostanie odrzucona to my tylko wiemy że wartości średnich Mi się różnią żeby dokonać mocniejszą analizę hipotezy  $H_1$  musimy wykonać kolejne testy -> analizę post-hoc

## **Sprawdź założenia modelu jednoczynnikowej analizy wariancji**

$H_0$  :  $\sigma_1^2 == \sigma_2^2 == \dots == \sigma_n^2$

$H_1$ :  $!H_0$

Można wykonać tylko jeden z poniższych testów ale w praktyce wykonuje się ich kilka i porównuje wyniki i jeśli są zgodne to jest okey.

### **1) Test Bartletta:**

Test ten jest najlepszy gdy jest rozkład normalny. Ale gdy normalności nie ma - są nawet drobne odstępstwa od niej - to ten test będzie bardzo złym wyborem. W takim przypadku wybieramy poniższe alternatywy - zaleca się test Levene'a

```
bartlett.test(response ~ dose, data = x)
```

## 2) Test Flingera-Killeena:

```
fligner.test(response ~ dose, data = x)
```

## 3) Test Levene'a:

Bardziej odporny na brak normalności - w takim przypadku daje bardziej wiarygodne wyniki. (gdy jest wykryta normalność - daje gorsze wyniki)

```
library(car)
leveneTest(response ~ dose, data = x)
```

## 4) Test Browna-Forsytha:

Modyfikacja testu Levene'a w której parametr położenia wyznaczany jest przez mediany a przez średnie.

```
library(car)
leveneTest(response ~ dose, data = x, center = "mean")
```

# ANALIZA POST-HOC:

Analiza po fakcie - wykonujemy ją jak główny test odrzuci  $H_0$ . Szukamy które grupy są do siebie równe a które się od siebie różnią. Stosujemy je aby zobaczyć gdzie znajdują się istotne różnice.

Bierzemy sobie (prawie) dozwolone pary, zestawiamy je sobie i je testujemy

$H_0: \mu_i = \mu_j$

$H_1: \mu_i \neq \mu_j$

## 1. Test t-studenta (parami)

tym sposobem rośnie nam błąd pierwszego rodzaju i żeby go kontrolować potrzebujemy skorygować p-wartości otrzymane - R robi to domyślnie korektą Holma. Szacuje wariancje tylko dla obserwacji określonych 2 grup.

```
pairwise.t.test(x$response, x$dose, data = x)
#orzytmujemy p-value dla każdego połączenia grup
```



## 2. Test HSD Turkeya:

Wykonuje oszacowanie wariancji na podstawie obserwacji  
WSZYTSKICH GRUP (Podobne własności do testu Fishera)

```
model_aov <- aov(response ~ dose, data = x)
TukeyHSD(model_aov)
#p-adj == p-value tableka opisana parami (ładniej opisane niż t-studenta)
```

```
#Inny test Turkeya - ładne podsumowanie na końcu - wskazuje grupy które są
#jednorodne, wskazują one gdzie nie ma istotnych różnic a gdzie są. Tam
#gdzie jest ta sama litera to między tymi grupami nie ma istotnych różnic.
library(agricolae)
model_aov <- aov(response ~ dose, data = x)
HSD.test(model_aov, "dose", console = TRUE)
```

## 3. Test Studenta-Newmana-Keulsa:

Test który może mieć większą moc, ale kosztem błędu 1.rodzaju - może go zawyżać. Ma najwięcej siły do wykrywania istotnych różnic.

```
#dzieli nam zmienne grupujące na dwie rozłączne grupy jednorodne.
#Elegancko odcina te grupy
model_aov <- aov(response ~ dose, data = x)
SNK.test(model_aov, "dose", console = TRUE)
```

## 4. Test LSD Fishera:

Podobne własności do testu Turkeya - podobne wyniki. Jest pomiędzy Student-Newmana-Keulsa a Schaffem.

```
model_aov <- aov(response ~ dose, data = x)
LSD.test(model_aov, "dose", p.adj = "holm", console = TRUE)
```

## 5. Test Scheffego:

Najbardziej konserwatywny test - bardzo ostrożny, najbardziej samozachowawczy - ma mniejszą moc ale dobrze kontroluje błędy 1. rodzaju.

```
model_aov <- aov(response ~ dose, data = x)
scheffe.test(model_aov, "dose", console = TRUE)
```

## ANALIZA KONTRASTÓW:

Rozszerzenie analizy post-hoc (bo post-hoc jest szczególnym przypadkiem analizy kontrastów)  
-> POMINIĘTE NA WYKŁADZIE

### Test Kruskala-Wallisa:

- Jeśli jest normalność - nie zaleca się stosowania tego testu bo jest on wtedy słabszy! Test główny - to test parametryczny - zakładał normalność. Ale nie zawsze musi być równość wariancji czy normalności - w takim przypadku będziemy korzystać z testu nieparametrycznego Kruskala-Wallisa:
  - nie zakłada on normalności i równości wariancji. Opiera się na rangach i jest rozszerzeniem testu Manna-whitneya-Wilcoxon.
- Tutaj dane badamy na podstawie mediany.

#### a. Agregacja danych:

```
aggregate(x$response, list(DOSE = x$dose),  
FUN = median)  
#posumowanie poprzez medianę - (lepiej tutaj operować na medianie)
```

#### b. Test główny:

```
kruskal.test(response ~ dose, data = x)
```

#### c. Testy post-hoc (gdy przyjmujemy $H_1$ ):

```
pairwise.wilcox.test(x$response, x$dose, data = x)  
#wykonywana parami
```

```
#inny test -> test DUNNA  
library(FSA)  
dunnTest(response ~ dose, data = x, method = "bh")
```