# Analiza wariancji

*Piotr J. Sobczyk*

*19 November 2016*

Zacznijmy zajęcia od klasycznego przykładu czyli testu Studenta dla dwóch prób.

$$x_{1,i} \sim N(\mu_1, \sigma^2),\ i = 1, ..., n_1$$
$$x_{2,i} \sim N(\mu_2, \sigma^2),\ i = 1, ..., n_2$$
$$n_1 + n_2 = n$$

Weźmy dane o wzroście piłkarzy i koszykarzy.

```
library(dplyr)
library(ggplot2)
wzrost=read.csv2("wzrost.csv")
attach(wzrost)
```

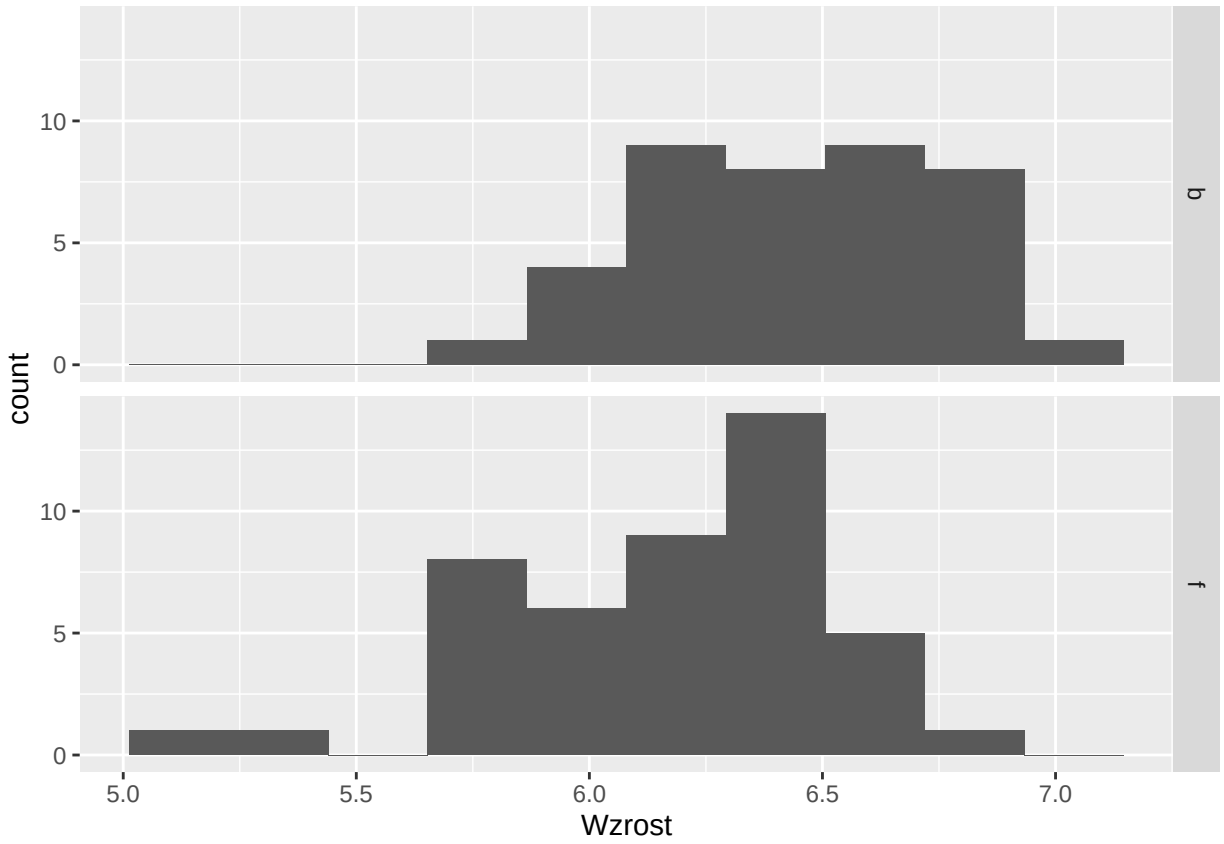
Sprawdzamy założenie normalności

```
wzrost %>%
  group_by(Dyscyplina) %>%
  summarise(shapiro.test(Wzrost)$p.value)
```

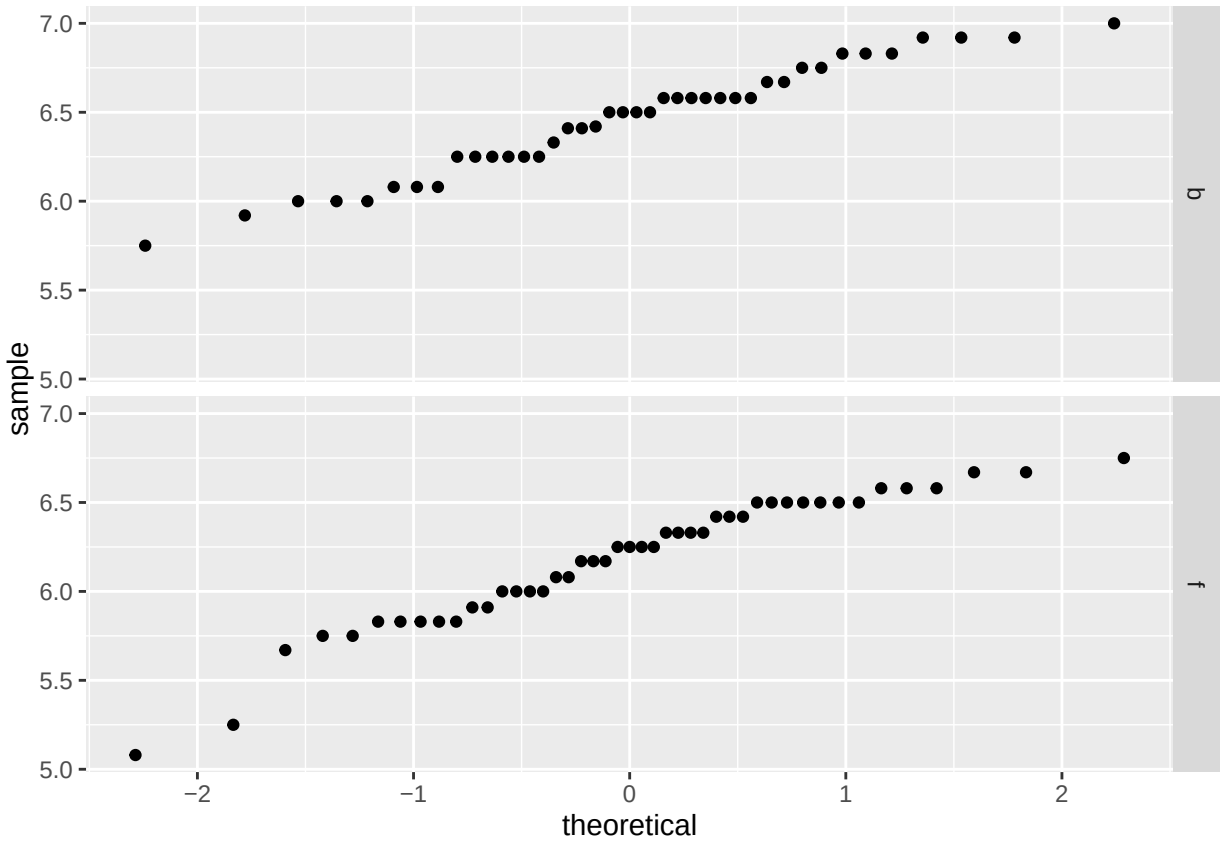
```
## # A tibble: 2 × 2
##   Dyscyplina `shapiro.test(Wzrost)$p.value`
##       <fctr>                         <dbl>
## 1          b                    0.31967963
## 2          f                    0.01609424
```

Jak wyglądają dane?

```
ggplot(wzrost) +
  geom_histogram(aes(x=Wzrost), bins = 10) +
  facet_grid(Dyscyplina~.)
```




```
ggplot(wzrost, aes(sample=Wzrost)) +
  stat_qq() +
  facet_grid(Dyscyplina~.)
```




Sprawdzenie założenia o równości wariancji

```
var.test(Wzrost~Dyscyplina)
```

```
##
##  F test to compare two variances
##
## data:  Wzrost by Dyscyplina
## F = 0.73668, num df = 39, denom df = 44, p-value = 0.3343
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3993554 1.3747946
## sample estimates:
## ratio of variances
##          0.7366753
```

I testowanie:

```
t.test(Wzrost~Dyscyplina)
```

```
##
##  Welch Two Sample t-test
##
## data:  Wzrost by Dyscyplina
## t = 3.7175, df = 82.907, p-value = 0.0003645
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1275687 0.4211535
## sample estimates:
```

```
## mean in group b mean in group f
##        6.453250        6.178889
```

```
t.test(Wzrost~Dyscyplina,alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  Wzrost by Dyscyplina
## t = 3.7175, df = 82.907, p-value = 0.0001822
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.1515952       Inf
## sample estimates:
## mean in group b mean in group f
##        6.453250        6.178889
```

```
t.test(Wzrost~Dyscyplina,alternative="greater",var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Wzrost by Dyscyplina
## t = 3.6841, df = 83, p-value = 0.0002039
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.1504837       Inf
## sample estimates:
## mean in group b mean in group f
##        6.453250        6.178889
```

Co zrobić gdybyśmy mieli dodaktowo baseballistów lub golfistów?

Porównania wielokrotne każda grupa przeciwko każdej? Problemem jest testowanie wielokrotne.

Pojedynczy t-test wykonywany jest na poziomie istotności $p$, równym np. 0.05. Ale przy wykonaniu serii testów nie możemy powiedzieć o wszystkich na raz, że p-stwo popełnienie błędu pierwszego rodzaju wynosi 0.05.

$$P(\text{odrzucenia jakiejkolwiek hipotezy} \mid \text{wszystkie są prawdziwe}) \leq \sum_{i}^{k} P(\text{odrzucenie i-tej hipotezy} \mid \text{i-ta hipoteza jest prawdz}$$

Już przy 10 testach (czyli 5 grupach) p-stwo błędu pierwszego rodzaju rośnie nam do 0.5! Do tego zagadnienia wrócimy później, ale zapamiętajmy póki co, że porównania wielokrotne rodzą problemy i wymagają od statystyka wyjątkowej uwagi.

### Analiza wariancji

Uprośćmy sobie zatem nieco nasz problem, albo inaczej, dokonajmy bardzo konkretnej generalizacji testu Studenta. Zamiast badać wszystkie różnice między średnimi, spróbujmy zbadać czy wszystkie średnie są sobie równe.

Towarzysyć nam będzie przykład wpływ analogów witaminy D na obecność antygenu CD14 u osób chorych na białaczkę. Dane pochodzą z pakietu **PBImisc**.

```
library(PBImisc)
data(AML)
?AML
summary(AML)
```

```
##     Mutation   CD14.control      CD14.D3        CD14.1906
##  CBFbeta:15   Min.   : 4.93   Min.   : 1.00   Min.   : 2.00
##  FLT3   :14   1st Qu.:29.93   1st Qu.:30.70   1st Qu.:31.14
##  None   :18   Median :47.05   Median :54.27   Median :47.46
##  Other  :19   Mean   :45.67   Mean   :50.33   Mean   :49.04
##               3rd Qu.:58.93   3rd Qu.:68.13   3rd Qu.:70.44
##               Max.   :94.35   Max.   :98.43   Max.   :92.83
##    CD14.2191
##  Min.   : 1.80
##  1st Qu.:24.66
##  Median :46.05
##  Mean   :46.30
##  3rd Qu.:63.90
##  Max.   :99.32
```

Przyglądnijmy się bliżej analogowi 2191

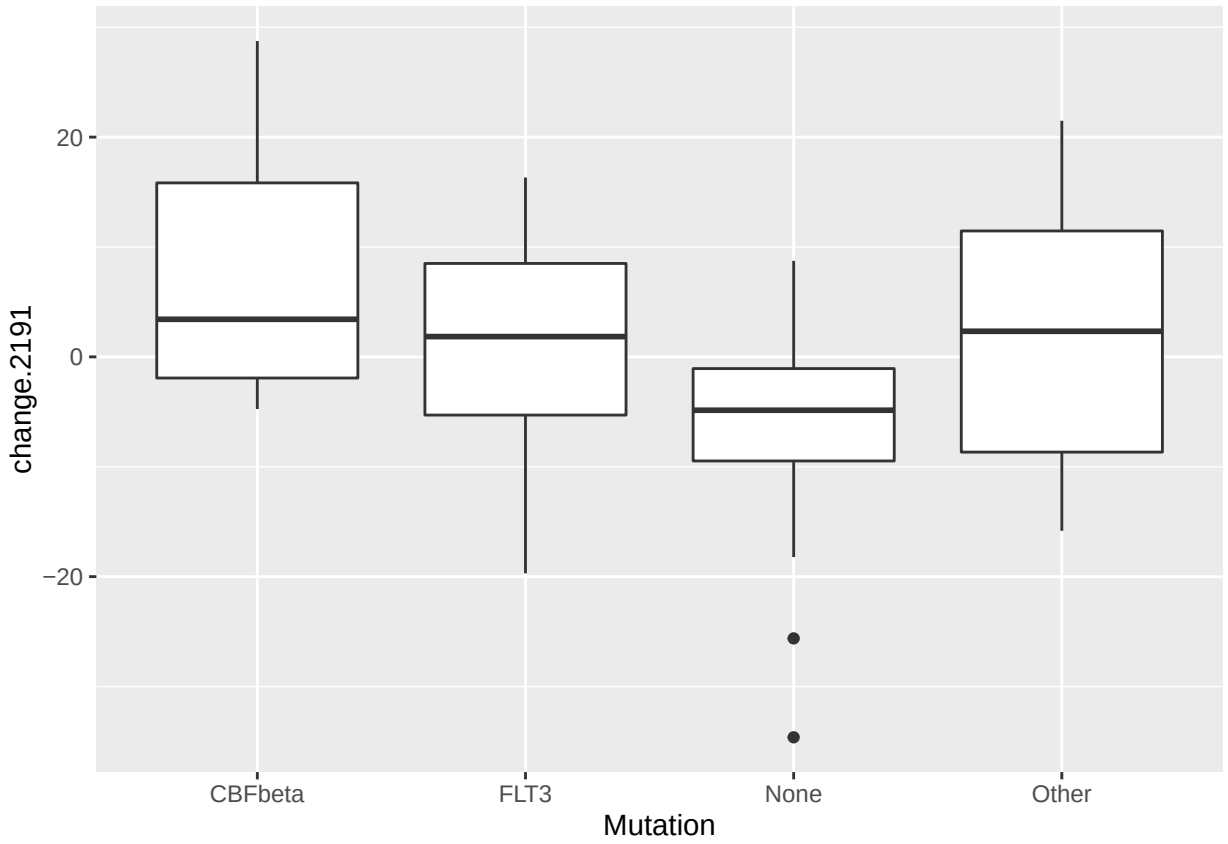
```
AML$change.2191=AML$CD14.2191-AML$CD14.control
```

Interesuje nas czy mutacja, wpływa na change.2191, a więc czy ten konkretny analogon jest bardziej odpowiedni dla którejś z mutacji

```
AML %>%
  group_by(Mutation) %>%
  summarise(mean(change.2191))
```

```
## # A tibble: 4 × 2
##   Mutation `mean(change.2191)`
##     <fctr>               <dbl>
## 1  CBFbeta           7.8140000
## 2     FLT3           0.8271429
## 3     None          -6.6805556
## 4    Other           1.7531579
```

```
ggplot(AML, aes(x=Mutation, y=change.2191)) +
  geom_boxplot()
```




A jak powiemy o tym w języku statystycznym? Mamy $k$ prób

$$x_{j,i} \sim N(\mu_j, \sigma^2),\ i = 1, ..., n_j\ ,\ j = 1, \ldots, kn_1 + \cdots + n_k = n$$

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \text{ vs. } H_1 : \exists_{l,m} \mu_l \neq \mu_m$$

```
aov.change=aov(change.2191~Mutation-1, data=AML)
summary(aov.change)
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## Mutation   4   1787   446.8   3.422 0.0136 *
## Residuals 62   8094   130.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To działa jak model liniowy! Każdy współczynnik odpowiada innemu poziomowi zmiennej objaśniającej. Żeby lepiej to zrozumieć popatrzmy na macierz planu $X$ w naszym modelu $y \sim X\beta + \epsilon$.

```
model.matrix(aov.change)[sample(1:66, 10),]
```

```
##    MutationCBFbeta MutationFLT3 MutationNone MutationOther
## 53               0            0            1             0
## 65               0            0            1             0
## 41               0            0            0             1
## 21               0            1            0             0
## 47               0            0            0             1
## 49               0            0            1             0
```

```
## 63               0            0            1            0
## 66               0            0            1            0
## 51               0            0            1            0
## 22               0            1            0            0
```

```
summary(lm(change.2191~Mutation, data=AML))
```

```
##
## Call:
## lm(formula = change.2191 ~ Mutation, data = AML)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.9394  -9.2113   0.3987   7.6766  20.9260
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.814      2.950   2.649 0.010236 *
## MutationFLT3    -6.987      4.246  -1.646 0.104924
## MutationNone   -14.495      3.995  -3.629 0.000578 ***
## MutationOther   -6.061      3.946  -1.536 0.129682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.43 on 62 degrees of freedom
## Multiple R-squared:  0.1787, Adjusted R-squared:  0.1389
## F-statistic: 4.495 on 3 and 62 DF,  p-value: 0.006427
```

Szczegoly na kursie z modeli liniowych, sprawdzamy czy są różnice w grupach. Jaka hipoteza zerowa?

```
summary(aov(change.2191~Mutation, data=AML))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Mutation     3   1761   586.9   4.495 0.00643 **
## Residuals   62   8094   130.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Jest wpływ czy go nie ma?

```
anova(lm(change.2191~Mutation, data=AML))
```

```
## Analysis of Variance Table
##
## Response: change.2191
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Mutation   3 1760.7  586.89  4.4955 0.006427 **
## Residuals 62 8094.1  130.55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(change.2191~Mutation, data=AML))
```

```
##
## Call:
## lm(formula = change.2191 ~ Mutation, data = AML)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -27.9394  -9.2113   0.3987   7.6766  20.9260
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.814      2.950   2.649 0.010236 *
## MutationFLT3    -6.987      4.246  -1.646 0.104924
## MutationNone   -14.495      3.995  -3.629 0.000578 ***
## MutationOther   -6.061      3.946  -1.536 0.129682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.43 on 62 degrees of freedom
## Multiple R-squared:  0.1787, Adjusted R-squared:  0.1389
## F-statistic: 4.495 on 3 and 62 DF,  p-value: 0.006427
```

Jak można te wyniki interpretować?

Trzeba sprawdzić, czy spelnione są zalozenia. Na szczęście założeń jest mniej niż przy modelu liniowym

```
library(car)
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
leveneTest(y = AML$change.2191, group = AML$Mutation)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.3931 0.7584
##       62
```

```
lm(change.2191~Mutation, data=AML) -> change.lm

shapiro.test(change.lm$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  change.lm$residuals
## W = 0.98394, p-value = 0.5513
```

Co dalej? Na jakie pytania możemy odpowiedzieć? Na przyklad czy srednia globalna to 0. Zauważmy, że poprzednio mieliśmy k+1 parametrów (k-klas + intercept), a tylko k-wartości do dopasowania. Dlatego R domyślnie uznał pierwszą klasę za poziom referencyjny.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = 0 \text{ vs. } H_1 : \exists_l \mu_l \neq 0$$

```
summary(aov(change.2191~Mutation-1, data=AML))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Mutation     4   1787   446.8   3.422 0.0136 *
## Residuals   62   8094   130.6
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### Różnice między średnimi

Wróćmy do zagadnienia testowanie różnic miedzy średnimi. Niech $H_{0,i,j} : \mu_i = \mu_j$

$$P(\forall_{i,j}\ H_{0,i,j} \text{ nie odrzucona gdy prawdziwa}) = 1 - P(\exists i,j\ H_{0,i,j} \text{ odrzucona gdy prawdziwa})$$

Rozpiszmy dalej:

$$P(\exists i,j\ H_{0,i,j} \text{ odrzucona gdy prawdziwa}) \leq \sum_{i,j} P(H_{0,i,j} \text{ odrzucona gdy prawdziwa})$$

Zatem,

$$P(\forall_{i,j}\ H_{0,i,j} \text{ nie odrzucona gdy prawdziwa}) \geq 1 - \sum_{i,j} P(H_{0,i,j} \text{ odrzucona gdy prawdziwa})$$

Jeśli będziemy testować pojedynczą hipotezę na poziomie ufności

$$1 - \frac{\alpha}{\binom{k}{2}},$$

to szansa, że nie odrzucimy żadnej prawdziwej hipotezy zerowej wynosi $1 - \alpha$.

Czyli jeśli chcemy być ostrożni, to trudniej nam będzie odrzucać hipotezy zerowe. Jeśli się nad tym zastanowić, to nie może być inaczej. Powyższe rozumownie nazywa się korektą Bonferoniego. Każdy test wykonujemy na poziomie istotności $\frac{\alpha}{m}$, gdzie $m$ jest liczbą testów. Są inne (lepsze) korekty: np. Sidaka, Holma. Ta ostatnia jest domyślnie stosowana przez R przy porównaniach wielokrotnych

```
pairwise.t.test(AML$change.2191, AML$Mutation)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  AML$change.2191 and AML$Mutation
##
##       CBFbeta FLT3   None
## FLT3  0.3148  -      -
## None  0.0035  0.2799 -
## Other 0.3148  0.8188 0.1420
##
## P value adjustment method: holm
```

Zauważmy na marginesie, że problem wielokrotnego testowania dotyczy także p-wartości dla współczynników w regresji liniowej. Jeśli chcemy wypowiadać się na temat wszystkich współczynników na raz, wykonujemy wielokrotne testowanie i powinniśmy wziąc odpowiednią korektę.

```
tmp=summary(lm(change.2191~Mutation-1, data=AML))
tmp$coefficients[,4]
```

```
## MutationCBFbeta    MutationFLT3    MutationNone   MutationOther
##      0.01023632      0.78739359      0.01584360      0.50609397
```

```
p.adjust(tmp$coefficients[,4])
```

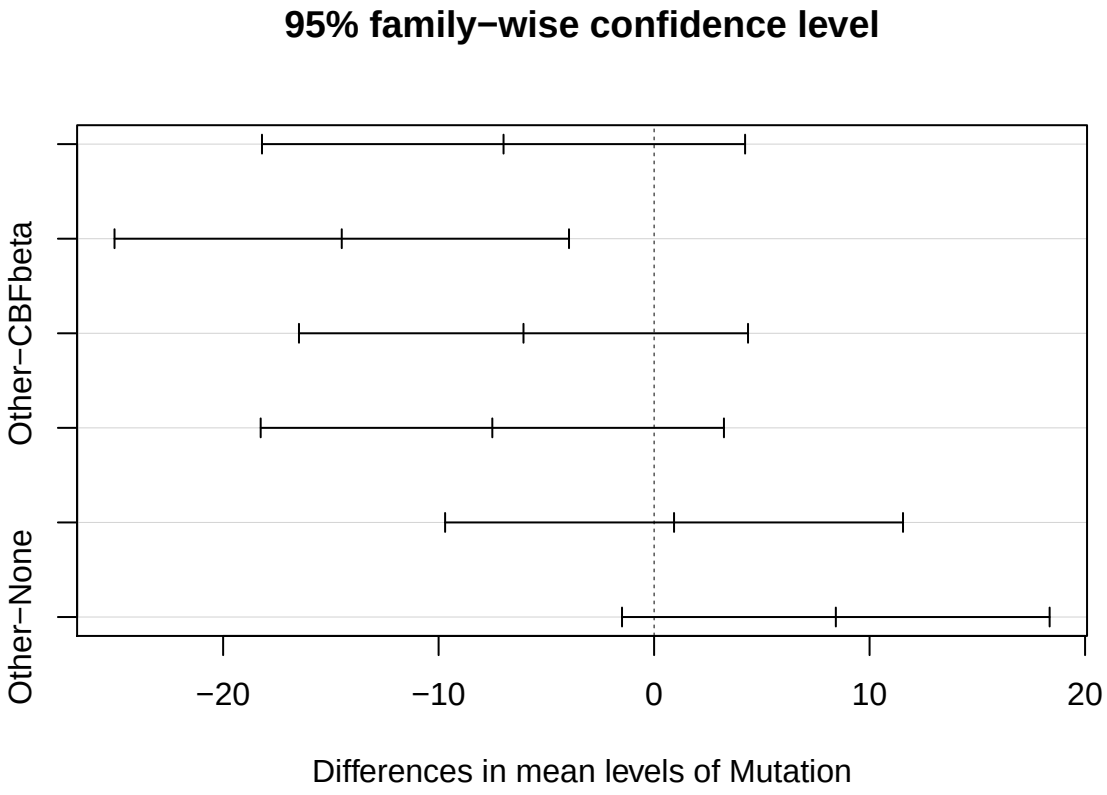
```
## MutationCBFbeta    MutationFLT3    MutationNone   MutationOther
##      0.04094528      1.00000000      0.04753080      1.00000000
```

Dla analizy wariancji możemy jednak zrobić trochę lepsze porównania. Jedna z metod nazywa się „testem uczciwych rzeczywistych różnic" (Honest Significant Differences) Tukeya.

```
TukeyHSD(aov(change.2191~Mutation-1, data=AML))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = change.2191 ~ Mutation - 1, data = AML)
##
## $Mutation
##                       diff        lwr       upr     p adj
## FLT3-CBFbeta    -6.986857 -18.196705  4.222991 0.3611805
## None-CBFbeta   -14.494556 -25.040500 -3.948611 0.0031652
## Other-CBFbeta   -6.060842 -16.479875  4.358191 0.4227192
## None-FLT3       -7.507698 -18.257120  3.241723 0.2629856
## Other-FLT3       0.926015  -9.698926 11.550956 0.9956627
## Other-None       8.433713  -1.488263 18.355690 0.1228184
```

```
plot(TukeyHSD(aov(change.2191~Mutation-1, data=AML)))
```




Na wykresie mamy przedziały ufności na różncie między grupami z uwzględnioną korektą na wielokrotne testowanie.

### Analiza wariancji wielokrotna

Co się dzieje kiedy mamy więcej niż jedną zmienną kategoryczną objaśniającą? Znowu szczegóły na kursie z modeli liniowych, ale zasada jest dokładnie taka sama.

$$y_{ijm} \sim N(\mu_{ij}, \sigma^2),\ i = 1, ..., k\ ,\ j = 1, \ldots, r,\ m = 1, \ldots, n_{ij}$$

Równoważnie,

$$y_{ijm} = \mu_{ij} + \epsilon_{ijm} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijm},\ i = 1, ..., k\ ,\ j = 1, \ldots, r,\ m = 1, \ldots, n_{ij}$$

Znowu liczba parametrów jest zbyt duża, w R domyślnie przyjmujemy:

$$\alpha_1 = 0,\ \beta_1 = 0,\ \forall_i \gamma_{1,i},\ \forall_j \gamma_{j,1}$$

Jakie hipotezy możemy testować?

- o braku interakcji $H_0 : \forall_{i,j} \gamma_{ij} = 0$ vs. $H_1 : \exists_{i,j} \gamma_{ij} \neq 0$
- o braku efektu pierwszej zmiennej $H_0 : \forall_i \alpha_i = 0$ vs. $H_1 : \exists_i alpha_i \neq 0$
- o braku efektu drugiej zmiennej $H_0 : \forall_j \beta_j = 0$ vs. $H_1 : \exists_j beta_j \neq 0$

### Przykład

Zbiór danych zawiera informacje o pacjentkach pewnego oddziału dla osób chorych psychiczne.

- **dr** to digit ratio - średnia arytmetyczna stosunku długości palca wskazującego do długości palca serdecznego w lewej i prawej dłoni.
- **c** to oznaczenie (grupy zaburzeń psychicznych)[https://pl.wikipedia.org/wiki/Zaburzenia_psychiczne#Podzia.C5.82_zaburze.C5.84_psychicznych_wed.C5.82ug_klasyfikacji_ICD-10]. A to zaburzenia F00-F29, B to zaburzenia F30-F99.
- **as** to rodzaj asymetrii twarzy pacjentek.

```
palce=read.table("palce.csv",sep=";",dec=",",head=TRUE)
attach(palce)
head(palce)
```

```
##     dr as ch
## 1 0.97  S  A
## 2 1.00  S  A
## 3 0.97 PA  B
## 4 1.02 PA  B
## 5 1.02 PA  A
## 6 0.96 LA  A
```

Jednoczynnikowa analiza wariancji raz jeszcze

```
m1=lm(dr~as)
head(model.matrix(m1))
```

```
##   (Intercept) asPA asS
## 1           1    0   1
## 2           1    0   1
## 3           1    1   0
## 4           1    1   0
## 5           1    1   0
## 6           1    0   0
```

```r
summary(m1)
```

```
##
## Call:
## lm(formula = dr ~ as)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.078889 -0.019474  0.002654  0.021111  0.070526
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.985217   0.006801 144.854   <2e-16 ***
## asPA        0.003671   0.010265   0.358    0.722
## asS         0.004256   0.010112   0.421    0.675
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03262 on 57 degrees of freedom
## Multiple R-squared:  0.003728,   Adjusted R-squared:  -0.03123
## F-statistic: 0.1066 on 2 and 57 DF,  p-value: 0.899
```

```r
sum(m1$residuals^2)
```

```
## [1] 0.06064643
```

```r
sum(lm(dr~1)$residuals^2)
```

```
## [1] 0.06087333
```

```r
anova(lm(dr~1),m1)
```

```
## Analysis of Variance Table
##
## Model 1: dr ~ 1
## Model 2: dr ~ as
##   Res.Df      RSS Df  Sum of Sq      F Pr(>F)
## 1     59 0.060873
## 2     57 0.060646  2 0.00022691 0.1066  0.899
```

```r
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: dr
##           Df   Sum Sq    Mean Sq F value Pr(>F)
## as         2 0.000227 0.00011345  0.1066  0.899
## Residuals 57 0.060646 0.00106397
```

```r
anova(lm(dr~ch))
```

```
## Analysis of Variance Table
##
## Response: dr
##           Df   Sum Sq    Mean Sq F value Pr(>F)
## ch         1 0.000934 0.00093444  0.9042 0.3456
## Residuals 58 0.059939 0.00103343
```

```r
summary(lm(dr~ch))
```

```
##
## Call:
## lm(formula = dr ~ ch)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08250 -0.02250 -0.00250  0.02556  0.07556
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.984444   0.005358 183.739   <2e-16 ***
## chB         0.008056   0.008471   0.951    0.346
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03215 on 58 degrees of freedom
## Multiple R-squared:  0.01535,    Adjusted R-squared:  -0.001626
## F-statistic: 0.9042 on 1 and 58 DF,  p-value: 0.3456
```

```r
t.test(dr~ch,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  dr by ch
## t = -0.9509, df = 58, p-value = 0.3456
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.025013058  0.008901947
## sample estimates:
## mean in group A mean in group B
##       0.9844444       0.9925000
```

### Dwuczynnikowa analiza wariancji

```r
m2=lm(dr~as+ch)
head(model.matrix(m2))
```

```
##   (Intercept) asPA asS chB
## 1           1    0   1   0
## 2           1    0   1   0
## 3           1    1   0   1
## 4           1    1   0   1
## 5           1    1   0   0
## 6           1    0   0   0
```

```r
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: dr
##           Df   Sum Sq    Mean Sq F value Pr(>F)
## as         2 0.000227 0.00011345  0.1071 0.8987
```

```
## ch         1 0.001303 0.00130300  1.2296 0.2722
## Residuals 56 0.059343 0.00105970
```

Można porównywać modele zagnieżdżone

```
anova(lm(dr~as),m2)
```

```
## Analysis of Variance Table
##
## Model 1: dr ~ as
## Model 2: dr ~ as + ch
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1     57 0.060646
## 2     56 0.059343  1  0.001303 1.2296 0.2722
```

```
anova(lm(dr~1),lm(dr~as))
```

```
## Analysis of Variance Table
##
## Model 1: dr ~ 1
## Model 2: dr ~ as
##   Res.Df      RSS Df  Sum of Sq      F Pr(>F)
## 1     59 0.060873
## 2     57 0.060646  2 0.00022691 0.1066  0.899
```

```
anova(lm(dr~as+ch))
```

```
## Analysis of Variance Table
##
## Response: dr
##           Df   Sum Sq    Mean Sq F value Pr(>F)
## as         2 0.000227 0.00011345  0.1071 0.8987
## ch         1 0.001303 0.00130300  1.2296 0.2722
## Residuals 56 0.059343 0.00105970
```

```
anova(lm(dr~1),lm(dr~as),lm(dr~as+ch))
```

```
## Analysis of Variance Table
##
## Model 1: dr ~ 1
## Model 2: dr ~ as
## Model 3: dr ~ as + ch
##   Res.Df      RSS Df  Sum of Sq      F Pr(>F)
## 1     59 0.060873
## 2     57 0.060646  2 0.00022691 0.1071 0.8987
## 3     56 0.059343  1 0.00130300 1.2296 0.2722
```

Za pomocą funkcji **anova** możemy porównywać jedynie modele zagnieżdżone

```
anova(lm(dr~ch),lm(dr~as)) # tak nie wolno!!!
```

```
## Analysis of Variance Table
##
## Model 1: dr ~ ch
## Model 2: dr ~ as
##   Res.Df      RSS Df   Sum of Sq F Pr(>F)
## 1     58 0.059939
## 2     57 0.060646  1 -0.00070754
```

Coś sie liczy, ale nie ma on sensu, nie możemy wnioskować na tej podstawie

Decyzję, o tym czy chcemy dodać, odjąc zmienną możemy podjąc na podstawie wartości kryterium informacyjnego (więcej na ten temat w kolejnych tygodniach zajęć).

```
anova(lm(dr~as),lm(dr~as+ch))
```

```
## Analysis of Variance Table
##
## Model 1: dr ~ as
## Model 2: dr ~ as + ch
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1     57 0.060646
## 2     56 0.059343  1  0.001303 1.2296 0.2722
```

```
anova(lm(dr~ch),lm(dr~as+ch))
```

```
## Analysis of Variance Table
##
## Model 1: dr ~ ch
## Model 2: dr ~ as + ch
##   Res.Df      RSS Df  Sum of Sq     F Pr(>F)
## 1     58 0.059939
## 2     56 0.059343  2 0.00059546 0.281 0.7561
```

```
drop1(lm(dr~as+ch),test="F")
```

```
## Single term deletions
##
## Model:
## dr ~ as + ch
##        Df  Sum of Sq      RSS     AIC F value Pr(>F)
## <none>               0.059343 -407.13
## as      2 0.00059546 0.059939 -410.53  0.2810 0.7561
## ch      1 0.00130300 0.060646 -407.82  1.2296 0.2722
```

```
anova(lm(dr~1),lm(dr~as))
```

```
## Analysis of Variance Table
##
## Model 1: dr ~ 1
## Model 2: dr ~ as
##   Res.Df      RSS Df  Sum of Sq      F Pr(>F)
## 1     59 0.060873
## 2     57 0.060646  2 0.00022691 0.1066  0.899
```

```
anova(lm(dr~1),lm(dr~ch))
```

```
## Analysis of Variance Table
##
## Model 1: dr ~ 1
## Model 2: dr ~ ch
##   Res.Df      RSS Df  Sum of Sq      F Pr(>F)
## 1     59 0.060873
## 2     58 0.059939  1 0.00093444 0.9042 0.3456
```

```
add1(lm(dr~1),scope=~as+ch,test="F")
```

```
## Single term additions
```

```
##
## Model:
## dr ~ 1
##        Df  Sum of Sq      RSS     AIC F value Pr(>F)
## <none>              0.060873 -411.60
## as      2 0.00022691 0.060646 -407.82  0.1066 0.8990
## ch      1 0.00093444 0.059939 -410.53  0.9042 0.3456
```

### Dwuczynnikowa analiza wariancji z interakcjami

```
m3=lm(dr~as*ch) # alternatywnie: lm(dr~as+ch+as:ch)
head(model.matrix(m3))
```

```
##   (Intercept) asPA asS chB asPA:chB asS:chB
## 1           1    0   1   0        0       0
## 2           1    0   1   0        0       0
## 3           1    1   0   1        1       0
## 4           1    1   0   1        1       0
## 5           1    1   0   0        0       0
## 6           1    0   0   0        0       0
```

```
drop1(m3,test="F")
```

```
## Single term deletions
##
## Model:
## dr ~ as * ch
##        Df  Sum of Sq      RSS     AIC F value Pr(>F)
## <none>              0.058931 -403.54
## as:ch   2 0.00041288 0.059343 -407.13  0.1892 0.8282
```

```
anova(lm(dr~as+ch),lm(dr~as*ch))
```

```
## Analysis of Variance Table
##
## Model 1: dr ~ as + ch
## Model 2: dr ~ as * ch
##   Res.Df      RSS Df  Sum of Sq      F Pr(>F)
## 1     56 0.059343
## 2     54 0.058931  2 0.00041288 0.1892 0.8282
```

```
step(m3)
```

```
## Start:  AIC=-403.54
## dr ~ as * ch
##
##         Df  Sum of Sq      RSS     AIC
## - as:ch  2 0.00041288 0.059343 -407.13
## <none>               0.058931 -403.54
##
## Step:  AIC=-407.13
## dr ~ as + ch
##
##        Df  Sum of Sq      RSS     AIC
## - as    2 0.00059546 0.059939 -410.53
## - ch    1 0.00130300 0.060646 -407.82
```

```
## <none>               0.059343 -407.13
##
## Step:  AIC=-410.53
## dr ~ ch
##
##        Df  Sum of Sq      RSS     AIC
## - ch    1 0.00093444 0.060873 -411.60
## <none>               0.059939 -410.53
##
## Step:  AIC=-411.6
## dr ~ 1
```

```
##
## Call:
## lm(formula = dr ~ 1)
##
## Coefficients:
## (Intercept)
##      0.9877
```

```
step(lm(dr~1),scope=~as*ch)
```

```
## Start:  AIC=-411.6
## dr ~ 1
##
##        Df  Sum of Sq      RSS     AIC
## <none>               0.060873 -411.60
## + ch    1 0.00093444 0.059939 -410.53
## + as    2 0.00022691 0.060646 -407.82
```

```
##
## Call:
## lm(formula = dr ~ 1)
##
## Coefficients:
## (Intercept)
##      0.9877
```

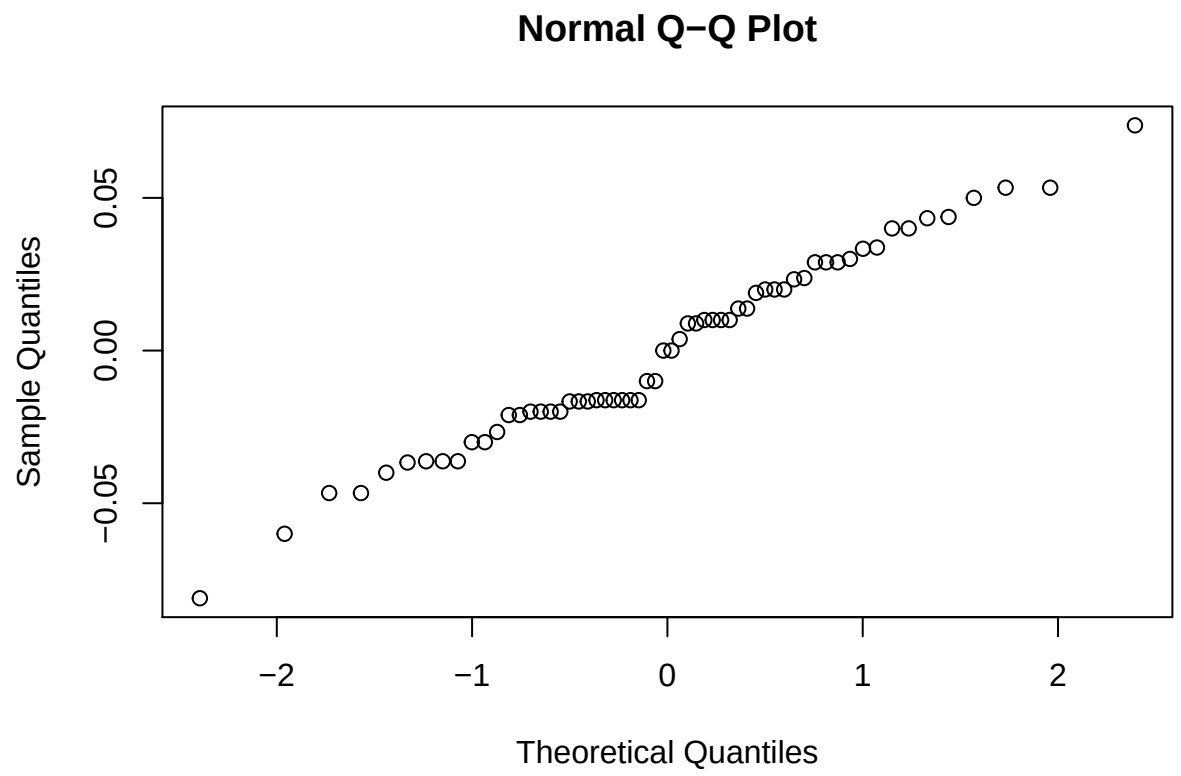
### Sprawdzanie założeń modelu m3

normalność

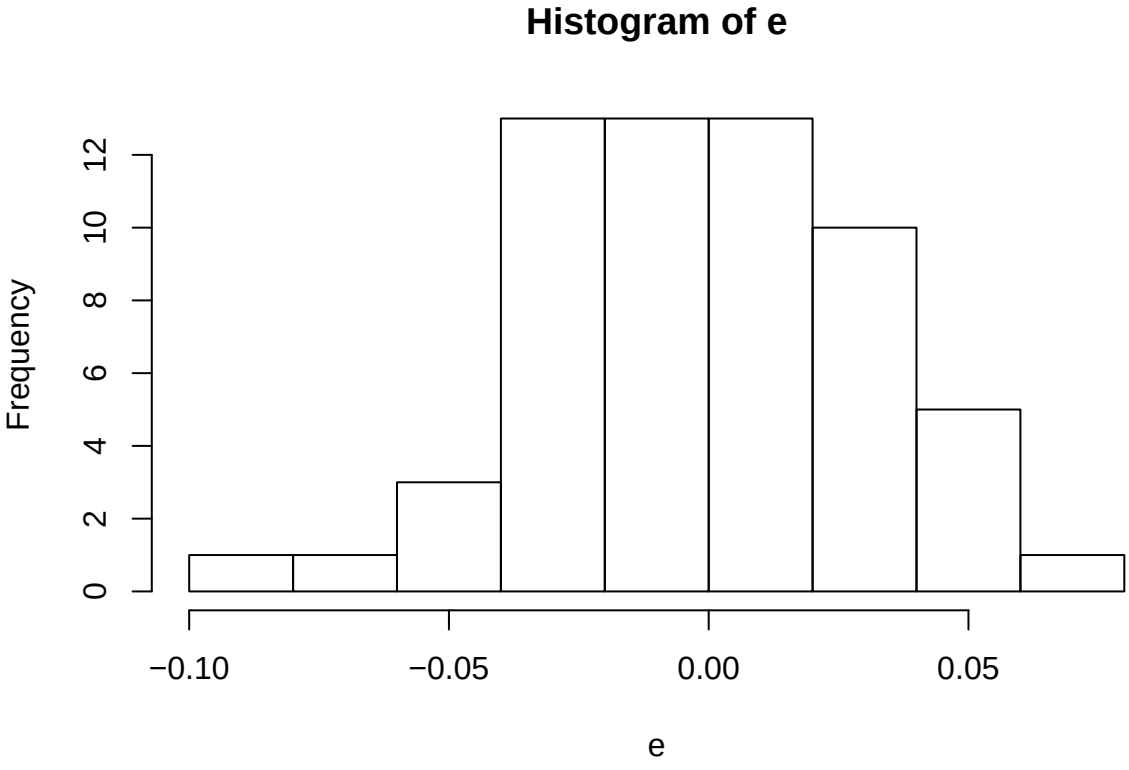
```
m3$residuals->e
shapiro.test(e)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  e
## W = 0.98207, p-value = 0.5219
```

```
qqnorm(e)
```




```
hist(e)
```




homoskedastyczność

```
bartlett.test(dr~as)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  dr by as
## Bartlett's K-squared = 1.2857, df = 2, p-value = 0.5258
```

```
bartlett.test(dr~ch)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  dr by ch
## Bartlett's K-squared = 0.33422, df = 1, p-value = 0.5632
```

```
bartlett.test(dr~paste(as,ch))
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  dr by paste(as, ch)
## Bartlett's K-squared = 4.2924, df = 5, p-value = 0.5081
```