Analiza wariancji

Na początek kilka ważnych faktów o analizie wariancji.

- Jest powszechnie stosowana metoda statystyczna pozwalająca na ocenę istotności różnic wielu średnich.
- Opracował ją i upowszechnił Ronald Fisher w latach 20 XX wieku.
- Jest to technika badania wyników (doświadczeń, obserwacji), które zależą od jednego (jednoczynnikowa analiza wariancji) lub kilku czynników działających równocześnie (dwuczynnikowa lub wieloczynnikowa analiza wariancji).
  - przykłady takich czynników: metody leczenia, płeć, sposoby nawożenia (Fisher!).
  - czynniki takie nazywamy *czynnikami grupującymi, klasyfikacyjnymi lub zmiennymi manipulacyjnymi.*
  - Każdy czynnik ma kilka poziomów lub wariantów (częstym wzorcem jest np. mało-średnio-dużo czegoś)
- Założenia analizy wariancji są podobne, jak w przypadku regresji liniowej - homogeniczność wariancji w grupach i normalność w grupach.

Model teoretyczny

Model strukturalny leżący u podstaw analizy wariancji ma następującą postać:

$$X_{ij} = \mu + \tau_j + \epsilon_{ij}$$

- X_{ij} - indywidualna obserwacja X numer i należąca do grupy j
- μ - ogólna średnia
- τ_j - efekt przynależenia do grupy j
- ϵ_{ij} - błąd

Układ hipotez przy analizie wariancji

Przy analizie wariancji mamy do czynienia z następującym układem hipotez:

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

W tym wypadku μ_1 do μ_k odnoszą się do średnich w populacji dla poszczególnych poziomów czynnika grupującego. Hipoteza zerowa stwierdza, że średnie te są sobie równe.

$$H_A : \neg H_0$$

Hipoteza alternatywna (ta, którą chcemy uzasadnić obalając hipotezę zerową!) głosi, że H_0 jest fałszem. Warto zwrócić uwagę, że H_0 jest sformułowana w bardzo specyficzny sposób i może okazać się fałszywa na wiele sposobów. Wszystkie średnie mogą być od siebie różne albo tylko niektóre. Sam test jednak powie nam jedynie tyle, że powinniśmy odrzucić hipotezę zerową. Czy możemy mimo wszystko dowiedzieć się czegoś na temat tego, w jaki sposób H_0 okazała się fałszywa? Tak! Żeby to zrobić, musimy przeprowadzić analizę *post-hoc*. O tym jednak powiemy sobie więcej na koniec tej części kursu (Uwaga! Możemy to zrobić wyłącznie wówczas, gdy ANOVA dała istotny statystycznie wynik).

Wariancja międzygrupowa i wewnątrzgrupowa

Logika analizy wariancji polega na porównywaniu wariancji *międzygrupowej* i *wewnątrzgrupowej*. Korzystamy z faktu, że – przy założeniu hipotezy zerowej – na podstawie obu możemy skonstruować estymator wariancji populacji. Jeśli hipoteza zerowa jest zaś fałszywa, to estymator „działa" tylko w jednym przypadku. Mechanizm ten będzie jaśniejszy, jeśli prześledzimy analizę wariancji na przykładzie.

Stwórzmy ramkę danych, w której mamy poziomów czynnika (jakiegokolwiek, tutaj użyhemy niewiele znaczacych poziomów „bardzo mało", „mało", „średnio", „dużo" oraz „bardzo dużo"), ale wszystkie wartości pochodzą z populacji o tych samych parametrach (takiej samej średniej oraz odchyleniu standardowym). Łatwo zauważyć, że jest to sytuacja, w której hipoteza zerowa jest prawdziwa.

```
# Nasz czynnik grupujący ma 5 poziomów, po 200 obserwacji każdy
faktor <- rep(c('bardzo mało', 'mało', 'średnio', 'dużo', 'bardzo dużo'), 200)
# Losujemy 1000 obserwacji z rozkładu o mu = 10 i sd = 5; po 200 na każdy poziom czynnika
wartosci <- rnorm(1000, mean = 10, sd = 5)
# Tworzymy z naszych wylosowanych obserwacji ramkę danych
df <- data.frame(condition = faktor, value = wartosci)
# Wyświetlamy pierwsze 6 wierszy
head(df)
```

```
##     condition     value
## 1 bardzo mało  4.514742
## 2        mało  9.184428
## 3     średnio  8.665288
## 4        dużo  9.434479
## 5 bardzo dużo 10.525129
## 6 bardzo mało  7.379373
```

Wariancja wewnętrz grup

Obliczmy wariancje w każdej z grup za pomocą funkcji `tapply`.

Krótkie przypomnienie jak działa `tapply`. Funkcja ta bierze wartości przekazane jako pierwszy argument (tutaj: `df$value`) i dzieli je według czynnika przekazanego jako drugi argument (tutaj: `df$condition`). Następnie do każdego elementu takiego podziału stosuje funkcję przekazaną jako trzeci argument – tutaj jest to `var` obliczające wariancję. Funkcja ta zwróci więc tyle wartości wariancji z próby (`var`) ile mieliśmy poziomów czynnika grupującego (`df$condition`) – w naszym wypadku 5.

```
tapply(df$value, df$condition, var)
```

```
## bardzo dużo bardzo mało        dużo        mało     średnio
##    24.61466    23.36752    28.24543    28.88171    25.06284
```

W następnym kroku obliczymy średnią z obliczonych właśnie wariancji. W ten sposób skonstruowaliśmy pierwszy z dwóch estymator wariancji populacji – obliczyliśmy wariancję w każdej z grup z osobna i wyciągnęliśmy z nich średnią.

```
mse <- mean(tapply(df$value, df$condition, var))
mse
```

```
## [1] 26.03443
```

Częściej spotkają się Państwo ze sposobem obliczania, w którym najpierw obliczamy sumę kwadratów a następnie średni kwadrat według następujących wzorów:

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

$$MSE = \frac{SSE}{n-k}$$

Sposób ten jest równoważny temu, co zrobiliśmy wcześniej, co można prześledzić wykonując umieszczony poniżej kod. W moim przekonaniu w tym sposobie nieco gorzej widać„o co chodzi' '.

```
# Dzielimy wyjściową ramkę danych według czynnika z kolumny `condition`
df_s <- split(df, df$condition)
# Obliczamy sumę kwadratów odchyleń
sse <- sum(sapply(df_s, function(x){sum((x$value - mean(x$value))^2)}))
sse
```

```
## [1] 25904.26
```

```
# Obliczamy średni kwadrat
mse <- sse/(1000-5)
mse
```

```
## [1] 26.03443
```

Wariancja między grupami

Nasz drugi estymator wariancji w populacji skonstruujemy, korzystajac z nabytej wcześniej wiedzy dotyczącej odchylenia standardowego średnich z próby. Innymi słowy - kolejny raz skorzystamy z Centralnego Twierdzenia Granicznego!

Wiemy z CTG, że:

$$\frac{\sigma^2}{n} = s^2_{\bar{X}}$$

Ten wzór możemy przekształcić w taki sposób, że uzyskamy:

$$\sigma^2 = n s^2_{\bar{X}}$$

Widzimy więc, że wariancja w populacji (σ^2) to po prostu wariancja średnich z próby ($s^2_{\bar{X}}$) pomnożona przez liczebność próby.

W takim razie, żeby wyestymować wariancję w populacji, powinniśmy obliczyć wariancję średnich z próby:

```
# Za pomocą `tapply` obliczamy średnią dla każdego poziomu `df$condition`
# Takich średnich w naszym wypadku otrzymujemy 5
# `var` oblicza wariancje dla tych 5 średnich
var(tapply(df$value, df$condition, mean))
```

```
## [1] 0.149931
```

```
# Żeby otrzymać estymator wariancji w populacji, musimy przemnożyć nasz wynik przez n = 200
# Bo tyle obserwacji mamy na każdy poziom czynnika grupującego
mstreat <- var(tapply(df$value, df$condition, mean))*200
mstreat
```

```
## [1] 29.98621
```

Częściej spotkają się Państwo ze sposobem obliczania, w którym najpierw obliczamy sumę kwadratów a następnie średni kwadrat według następujących wzorów:

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

$$MSA = \frac{SSA}{k-1}$$

```
ssa <- (sum((tapply(df$value, df$condition, mean) - mean(df$value))^2))*200
ssa
```

```
## [1] 119.9448
```

```
msa <- ssa/(5-1)
msa
```

```
## [1] 29.98621
```

Zwróćmy uwagę, że obliczona przez nas wartość będzie dobrym estymatorem wariancji w populacji wyłącznie wtedy, gdy próby pochodzą z tej samej populacji, ponieważ tylko takiej sytuacji dotyczy przywołane sformułowanie Centralnego Twierdzenia Granicznego! W innym wypadku tak obliczony „estymator'' nie będzie działać i ten fakt wykorzystamy do skonstruowania testu statystycznego!

Statystyka testowa

Statystyka testowa F będzie proporcją między dwiema obliczonymi własnie wartościami. W przypadku, kiedy grupa nie ma znaczenia (czyli, biorac pod uwagę nasz model teoretyczny, $\theta = 0$), spodziewamy się, że statystyka ta będzie wynosić 1 (taka jest jej wartość oczekiwana). Dlaczego? Ponieważ obie wartości są estymatorami tego samego parametru!

$$F = \frac{MSA}{MSE}$$

```
f_stat <- mstreat/mse
f_stat
```

```
## [1] 1.15179
```

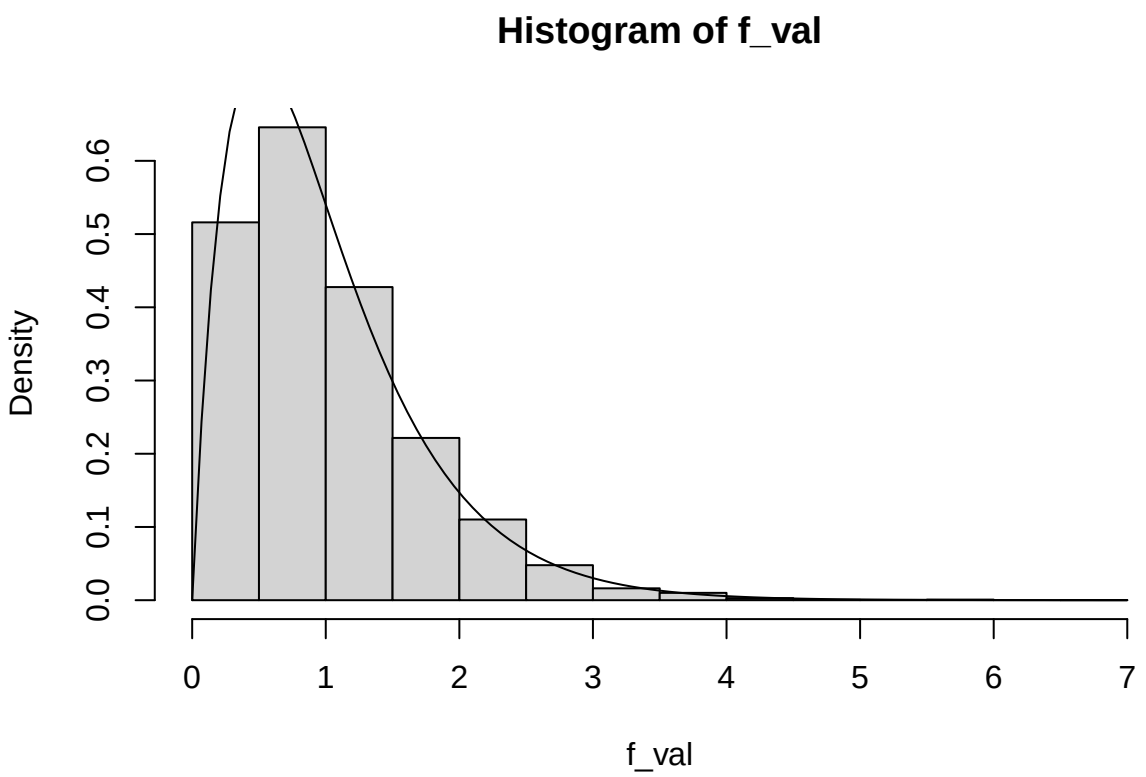
Czy wielkośc statystyki testowej, którą otrzymaliśmy, jest duża czy mała? Możemy zobaczyć, jaki rozkład ma interesująca nas statystyka testowa, przeprowadzając odpowiednią symulację.

Możemy również posłużyc się rozkładem teoretycznym F. Ma on dwa parametry - liczbę stopni swobody licznika (czyli $k-1$, gdzie k to liczba czynników) oraz liczbę stopni swobody licznika ($n-k$, gdzie k to liczba czynników a n to całkowita liczebność próby).

Na początek symulacja. Sprawdzamy, jaki rozkład ma statystyka testowa F przy założeniu prawdziwości hipotezy zerowej. W tym celu powtórzymy nasz eksperyment 10 000 razy i obliczamy dla każdego powtórzenia statystykę F. Następnie nakładamy na histogram statystyk F uzyskanych w symulacji odpowiednią krzywą z teoretycznego rozkładu F. Finalnie możemy zobaczyć, że histogram z symulowanymi wartościami F oraz krzywa rozkładu teoretycznego świetnie do siebie pasują.

```
f_val <- replicate(10000,
                   {
                     # Najpierw symulujemy naszą próbę
                     faktor = rep(c('bardzo mało', 'mało', 'średnio', 'dużo', 'bardzo dużo'),
                                  200)
                     wartosci <- rnorm(1000, mean = 10, sd = 5)
                     df <- data.frame(condition = faktor, value = wartosci)
                     # Obliczamy MSE i MS_Treatment
                     mse <- mean(tapply(df$value, df$condition, var))
                     mstreat <- var(tapply(df$value, df$condition, mean)) * 200
                     # Obliczamy statystykę testową F czyli iloraz tych dwóch wartości
                     mstreat / mse
```

```
                    })
# Wyniki symulacji możemy przedstawić na histogramie
hist(f_val, freq = FALSE)
curve(df(x, df1 = 5-1, df2 = 995), add = TRUE)
```



Skoro obliczyliśmy już statystykę testową F, to musimy jeszcze sprawdzić, czy uzyskanie takiej lub bardziej skrajnej wartości jest mniej prawdopodobne niż 5% (przyjmijmy taki konwencjonalny poziom α). Żeby to zrobić, musimy posłużyć się dystrybuantą rozkładu F, czyli funkcją `pf`. Rozkład F ma dwa parametry (dwie liczby stopni swobody). `df1` będzie tutaj wynosiła $k-1$ gdzie k jest liczbą poziomów czynnika grupującego, a `df2` będzie wynosiło $n-k$ gdzie n to (całkowita) liczebność próby.

```
pf(f_stat, df1 = 5-1, df2 = 995, lower.tail = FALSE)
```

```
## [1] 0.3307223
```

Obliczyliśmy właśnie p-wartość i tym samym przeprowadziliśmy naszą pierwszą analizę wariancji. Oczywiście przeprowadzając analizę wariancji nie musimy wykonywać ręcznie obliczeń. Jak zwykle możemy się posłużyć wbudowaną funkcją R, którą w tym wypadku jest `aov`.

```
summary(aov(value ~ condition, df))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## condition     4    120   29.99   1.152  0.331
## Residuals   995  25904   26.03
```

Analiza wariancji na przykładzie

Prześledźmy analizę wariancji na przykładzie. W bibliotece standardowej R znajduje się zbiór danych `PlantGrowth`, dotyczący wpływu pewnych czynników na obfitość plonów. W jednej kolumnie mamy wagę roślin, w drugiej to, do jakiej grupy przynależy dana obserwercja. Grupy były trzy - dwie eksperymentalne (`trt1` oraz `trt2`) i oraz grupa kontrolna (`ctrl`). Za pomocą analizy wariancji możemy zbadać, czy warunki eksperymentalne miały wpływ na wagę roślin.

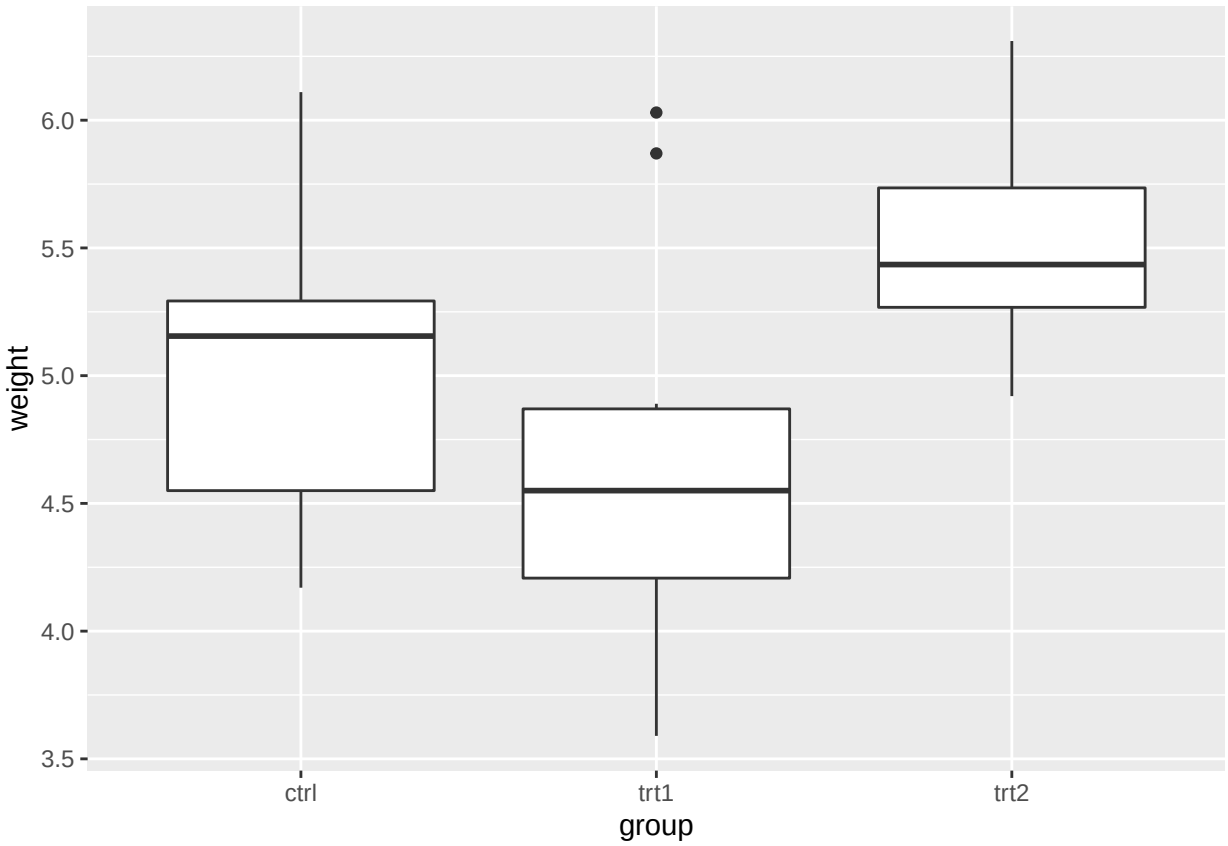
Nasza hipoteza zerowa będzie stwierdzać, że średnia waga jest taka sama we wszystkich grupach. Innymi słowy, że obserwacje dla wszystkich trzech poziomów czynnika grupującego pochodzą z populacji o tym samym parametrze średniej. Hipoteza alterantywna będzie zaś temu zaprzeczać. Oto zbiór danych (proszę wybaczyc, że wydrukuje go w całości, ale jest dość krótki):

```
PlantGrowth
```

```
##    weight group
## 1    4.17  ctrl
## 2    5.58  ctrl
## 3    5.18  ctrl
## 4    6.11  ctrl
## 5    4.50  ctrl
## 6    4.61  ctrl
## 7    5.17  ctrl
## 8    4.53  ctrl
## 9    5.33  ctrl
## 10   5.14  ctrl
## 11   4.81  trt1
## 12   4.17  trt1
## 13   4.41  trt1
## 14   3.59  trt1
## 15   5.87  trt1
## 16   3.83  trt1
## 17   6.03  trt1
## 18   4.89  trt1
## 19   4.32  trt1
## 20   4.69  trt1
## 21   6.31  trt2
## 22   5.12  trt2
## 23   5.54  trt2
## 24   5.50  trt2
## 25   5.37  trt2
## 26   5.29  trt2
## 27   4.92  trt2
## 28   6.15  trt2
## 29   5.80  trt2
## 30   5.26  trt2
```

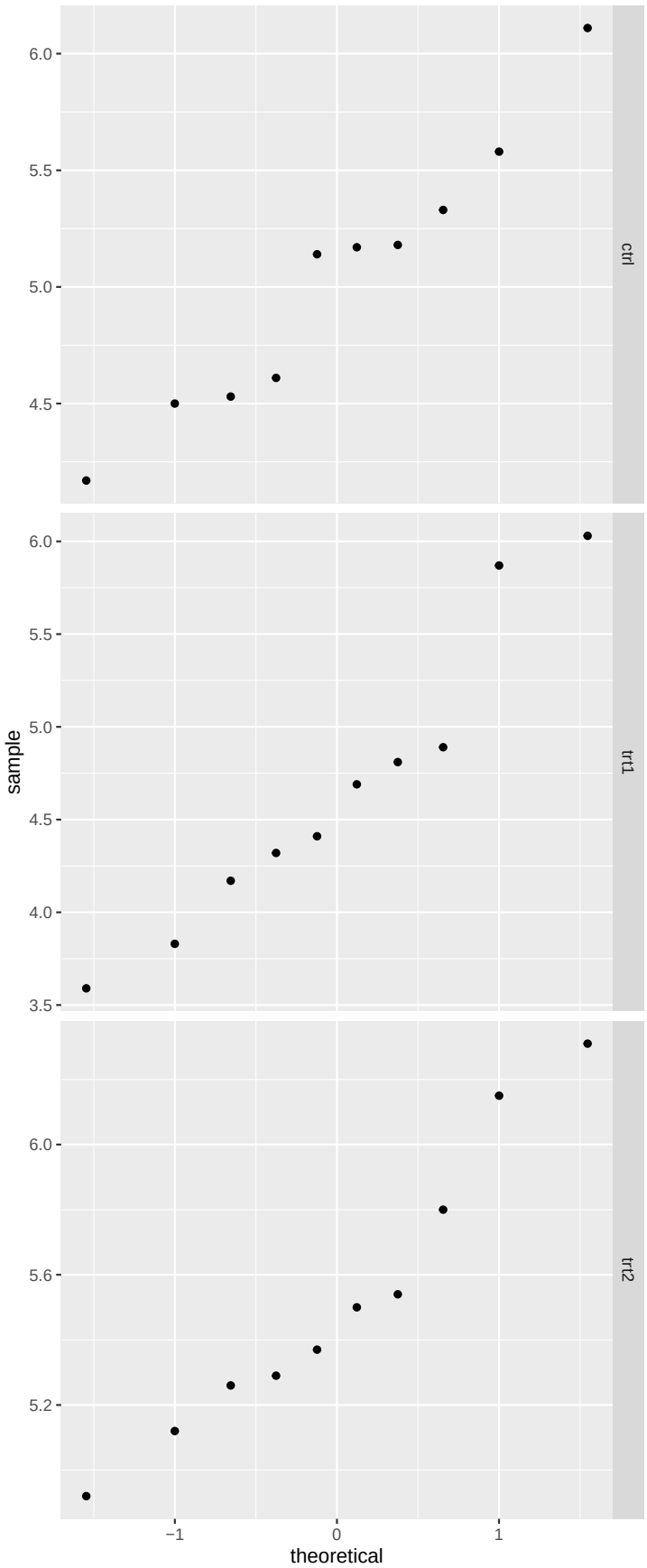
Na początek zawsze warto stworzyć wizualizację danych - między innymi po to, aby skontrolować to, czy nasze dane spełniają założenia analizy wariancji. Zasadniczo założenia są, podobnie jak przy regresji, dwa - równa wariancja oraz normalność w grupach. Na podstawie wykresu pudełkowego możemy "na oko" ocenić, czy mamy prawo zakładać, że założenia są spełnione. Te dane nie wyglądają bardzo źle – nie mamy powodów by powiedzieć, że w sposób oczywisty dane nie spełniają założeń.

```
library(ggplot2)
qplot(x = group, y = weight, geom = 'boxplot', data = PlantGrowth)
```



Normalność w grupach możemy "na oko" ocenić za pomocą wykresu kwantyl-kwantyl dla każdej grupy z osobna. Przy małych próbach może to okazać się to bardziej pomocne niż formalne testy na normalność rozkładu. Nie chcemy, żeby w sposób wyraźny i podpadający pod jakiś wzór odbiegały od przekątnej.

```
p <- ggplot(PlantGrowth, aes(sample = weight))
# scales = "free" sprawia, że skale na każdym wykresie są osobne.
p +
  stat_qq() +
  facet_grid(group~., scales = 'free')
```


6.0
5.5
5.0
4.5
ctrl
6.0
5.5
5.0
4.5
4.0
3.5
trt1
sample
6.0
5.6
5.2
trt2
−1
0
1
theoretical

Następnie możemy "ręcznie" przeprowadzić analizę wariancji w R. Najpierw musimy policzyć sumę kwadratów między grupami i wewnatrz grup, następnie obliczyć średni kwadrat w obu przypadkach. Po obliczeniu tych dwóch rzeczy możemy obliczyć statystykę testową F, dzieląc jedną przez drugą, i sprawdzić, korzystając z dystrybuanty rozkładu F, czy odrzucamy hipotezę zerową.

```
df_s <- split(PlantGrowth, PlantGrowth$group)

sse <- sum(sapply(df_s, function(x){sum((x$weight - mean(x$weight))^2)}))
print('SSE')
## [1] "SSE"
sse
## [1] 10.49209

mse <- sse/(nrow(PlantGrowth) - (length(levels(PlantGrowth$group))))
print('MSE')
## [1] "MSE"
mse
## [1] 0.3885959


ssa <-(sum((tapply(PlantGrowth$weight, PlantGrowth$group, mean) -
           mean(PlantGrowth$weight))^2))*(nrow(PlantGrowth)/3)
print('SSA')
## [1] "SSA"
ssa
## [1] 3.76634

msa <- ssa/(length(levels(PlantGrowth$group))-1)
print('MSA')
## [1] "MSA"
msa
## [1] 1.88317

f_stat <- msa/mse

print('Statystyka testowa F')
## [1] "Statystyka testowa F"
f_stat
## [1] 4.846088

pval <- pf(f_stat, (length(levels(PlantGrowth$group))-1), (nrow(PlantGrowth) - (length(levels(PlantGrow
          lower.tail = FALSE)
print('Wartość p')
## [1] "Wartość p"
pval
## [1] 0.01590996
```

To samo możemy wykonać za pomoca wbudowanej funkcji `aov`. Jak widzimy, uzyskane w eksperymencie rezultaty uprawniają nas do odrzucenia hipotezy zerowej. Możemy więc stiwerdzić, że proby nie pochodzą z populacji o takich samych średnich.

```
fit <- aov(weight ~ group, data = PlantGrowth)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
```

```
## group        2  3.766  1.8832   4.846 0.0159 *
## Residuals   27 10.492  0.3886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testy *post-hoc*

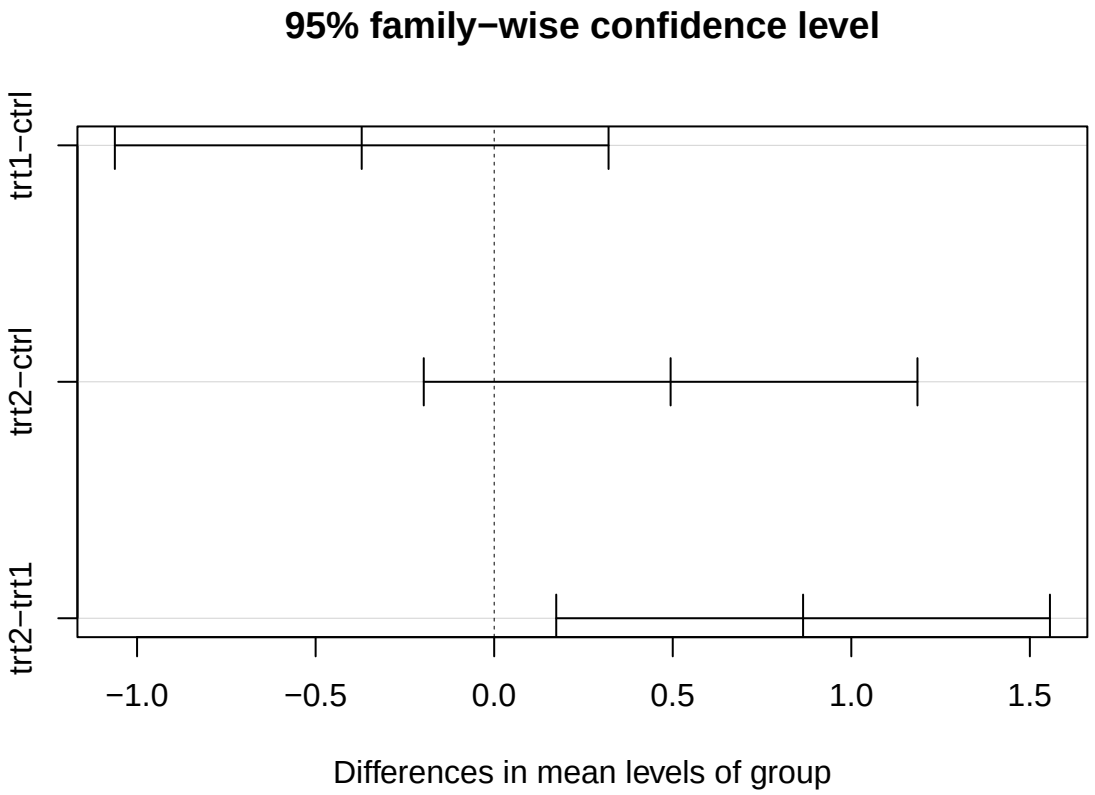
Często nie tylko interesuje nas fakt, że odrzuciliśmy hipotezę, zgodnie z którą wszystkie próby pochodzą z populacji o tej samej średniej, ale chcielibyśmy móc ustalić, między którymi grupami występują statytycznie istotne różnice. Tutaj przydatne okazują się tzw. testy *post-hoc*. Możemy ich użyć tylko jeśli nasza główna metoda (tutaj – ANOVA) dała statystycznie istotny wynik. Zaletą testów *post-hoc* jest to, że uwzględniają fakt, że są *post-hoc* i podają bardziej wiarygodne wyniki niż np. seria testów *t*. Testów *post-hoc*, które możemy zastosować przeprowadziwszy analizę wariancji jest kilka, najbardziej popularnym jest chyba test HSD (*honest significant difference*) Tukeya. Możemy taki test łatwo przeprowadzić za pomocą wbudowanej w R funkcji `TukeyHSD`. Na poniższym wydruku widzimy, że statystycznie istotna jest różnica pomiędzy dwoma grupami eksperymentalnymi (ale już nie między każdą z nich a grupą kontrolną!).

```
TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight ~ group, data = PlantGrowth)
##
## $group
##             diff        lwr       upr     p adj
## trt1-ctrl -0.371 -1.0622161 0.3202161 0.3908711
## trt2-ctrl  0.494 -0.1972161 1.1852161 0.1979960
## trt2-trt1  0.865  0.1737839 1.5562161 0.0120064
```

Możemy wyniki naszej analizy przedstawić w formie graficznej za pomocą funkcji `plot`. Na osi Y mamy zestawienia ze sobą poziomów czynnika (tutaj - `group`), na osi X zaś 95% przedziały ufności różnicy średnich między nimi. Istotne statystycznie są te różnice, które nie obejmują zera (proszę przypomnieć sobie testowanie hipotez za pomocą przedziałów ufności).

```
plot(TukeyHSD(fit))
```



Wielkość efektu

W przypadku analizy wariancji standardowo używa się miary η^2. Oblicza się ją dzieląc międzygrupową sumę kwadratów przez całkowitą sumę kwadratów.

```
# Całkowita suma kwadratów
sst <- sum((PlantGrowth$weight - mean(PlantGrowth$weight)) ** 2)
eta_kw = ssa/sst
eta_kw
```

```
## [1] 0.2641483
```

Interpretacja tej miary jest analogiczna jak w przypadku znanego już Państu R^2 i jest to stosunek zmienności wyjaśnianej przez nasz predyktor do całkowitej zmienności zmiennej zależnej. Jeżeli mamy taką ochotę, to mozemy skorzystać z funkcji obliczającą tę miarę w jednej z wielu bibliotek, które taką funkcję oferują. Ja skorzystałem tutaj z `DescTools` i funkcji `EtaSq` (ale jest ich wiele!).

```
DescTools::EtaSq(fit)
```

```
##          eta.sq eta.sq.part
## group 0.2641483   0.2641483
```

Analiza wariancji a model liniowy

Istnieją bardzo podobne podobieństwa między analizą wariancji i modelem liniowym. Można nawet stwierdzić, że analiza wariancji to po prostu szczególny przypadek modelu liniowego. Fakt, że omawiane są oddzielnie podyktowane jest w znacznej mierze zaszłościami historycznymi. W artykule dostępnym w sekcji dla zalogowanych "Multiple Regression As A General Data-Analytic System" autorstwa Jacoba Cohena znajduje się dobre wyjaśnienie wielu zawiłości z tym związanych.

Zobaczmy sobie podsumowanie jednoczynnikowej analizy wariancji oraz modelu liniowego z jednym kategorialnym predyktorem.

Analiza wariancji:

```
fit_aov <- aov(weight ~ group, data = PlantGrowth)
summary(fit_aov)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## group        2  3.766  1.8832   4.846 0.0159 *
## Residuals   27 10.492  0.3886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
DescTools::EtaSq(fit_aov)
```

```
##          eta.sq eta.sq.part
## group 0.2641483   0.2641483
```

Model liniowy:

```
fit_lm <- lm(weight ~ group, data = PlantGrowth)
summary(fit_lm)
```

```
##
## Call:
## lm(formula = weight ~ group, data = PlantGrowth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0710 -0.4180 -0.0060  0.2627  1.3690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0320     0.1971  25.527   <2e-16 ***
## grouptrt1    -0.3710     0.2788  -1.331   0.1944
## grouptrt2     0.4940     0.2788   1.772   0.0877 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6234 on 27 degrees of freedom
## Multiple R-squared:  0.2641, Adjusted R-squared:  0.2096
## F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591
```

Zwróćmy uwagę na kilka rzeczy. Dla analizy wariancji hipoteza zerowa wygląda tak:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

Dla modelu liniowego hipotezą zerową było:

$$H_0 : R^2 = 0$$

Hipotezy te, chociaż pozornie brzmiące różnie, w rzeczywistości są sobie matematycznie równoważne i można je przetestować tym samym testem (na co wskazuje nam w wydruku R taka sama wartość statystyki testowej F). Możemy to zilustrować takim oto rozumowaniem.

Jeżeli hipoteza zerowa analizy wariancji jest prawdziwa (czyli średnie w grupach są w populacji identyczne), to znajomość tego, do której grupy dana obserwacja należy, nie pozwala nam przewidywać wartości tej obserwacji lepiej niż po prostu znajomość ogólnej średniej. W takiej sytuacji nasz model nie wyjaśnia żadnej części wariancji zmiennej zależnej czyli R^2 jest równe zeru. Warto również zwrócić uwagę na to, że R^2 jest równy obliczonemu przez nas η^2, co nie powinno dziwić biorać pod uwagę sposób w jaki oblicza się obie te wartości.

Na kolejnych zajęciach, gdy będziemy zajmować się nieco bardziej skomplikowanymi wariantami analizy wariancji (w których uwzględniamy więcej czynników), zobaczymy w jaki sposób tę paralelę między modelem liniowym a analizą wariancji można pociągnąć dalej.