A tutorial on ANOVA and Kruskal-Wallis test in R

Version 1.0.0

Konstantinos I. Bougioukas

04/11/2021

Contents

Objectives

- Applying hypothesis testing
- Compare more than two independent samples
- Post-hoc analysis
- Interpret the results

We will need to download and load the following packages for the notes:

```
library(ggpubr)
library(see)


library(rstatix)
library(EnvStats)
library(here)
library(tidyverse)
```

1 Introduction

The one-way analysis of variance (one-way ANOVA) or the non-parametric Kruskal-Wallis test are used to detect whether there are any differences between more than two independent (unrelated) samples.

Althought, these tests can detect a difference between several groups they do not inform about which groups are different from the others. At first sight we might clarify the question by comparing all groups in pairs with t-tests or Wilcoxon-Mann-Whitney (WMW) tests. However, that procedure may lead us to the wrong conclusions (known as multiple comparisons problem). Why is this procedure inappropriate? Quite simply, because we would be wrongly testing the null hypothesis. Each comparison one conducts increases the likelihood of committing at least one Type I error within a set of comparisons (famillywise Type I error rate).

This is the reason why, after an ANOVA or Kruskal-Wallis test concluding on a difference between groups, we should not just compare all possible pairs of groups with t-tests or WMW. Instead we perform statistical tests that take into account the number of planned comparisons (post hoc tests). Some of the more commonly used ones are Tukey's test, Dunn's test and Bonferroni correction.

2 One-way Analysis of Variance (one-way ANOVA)

One-way analysis of variance, usually referred to as one-way ANOVA, is a statistical test used when we want to compare several means. We may think of it as an extension of Student's t-test to the case of more than two samples.

2.1 Research question

Consider the example of the variations between weight loss according to four different types of diet. The question that may be asked is: does the average weight loss differ according to the diet?

2.2 H0 and H1 Hypotheses

- H_0: all group means are equal (the means of weight loss in the four diets are equal)
- H_1: at least one group mean differs from the others (there is at least one diet with mean weight loss different from the others)

2.3 Preraring the data

We import the data:

```r
library(readxl)
dataDWL <- read_excel(here("data", "dataDWL.xlsx"), col_names=TRUE)
dataDWL
```

Table 1: Diet Weight Loss Data (first and last 5 rows)

WeightLoss	Diet
9.9	A
9.6	A
8	A
4.9	A
10.2	A
...	NA
11.8	D
7.1	D
9.4	D
13.7	D
13.7	D

We inspect the data:

```
glimpse(dataDWL)
```

```
## Rows: 60
## Columns: 2
## $ WeightLoss <dbl> 9.9, 9.6, 8.0, 4.9, 10.2, 9.0, 9.8, 10.8, 6.2, 8.3, 12~
## $ Diet       <chr> "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A",~
```

The dataset `dataDWL` has 60 participants and includes two variables. The numeric `WeightLoss` variable and the `Diet` variable (with levels “A”, “B”, “C” and “D”) which should be converted from character to a factor variable using the `factor()`:

```
dataDWL <- dataDWL %>%
  mutate(Diet = factor(Diet))
glimpse(dataDWL)
```

```
## Rows: 60
## Columns: 2
## $ WeightLoss <dbl> 9.9, 9.6, 8.0, 4.9, 10.2, 9.0, 9.8, 10.8, 6.2, 8.3, 12~
## $ Diet       <fct> A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, B, B, B, ~
```

2.4 Assumptions

1. The dependent variable should be approximately **normally** distributed for all groups
2. The data in groups have similar **variance** (homoscedasticity)

The assumptions can be checked visually using **dot plots and histograms**. Moreover, summary statistics (means, medians), significance tests available to check for normality (e.g., Shapiro-Wilk test) and for equality of variances (e.g., Levene's test) can be used.

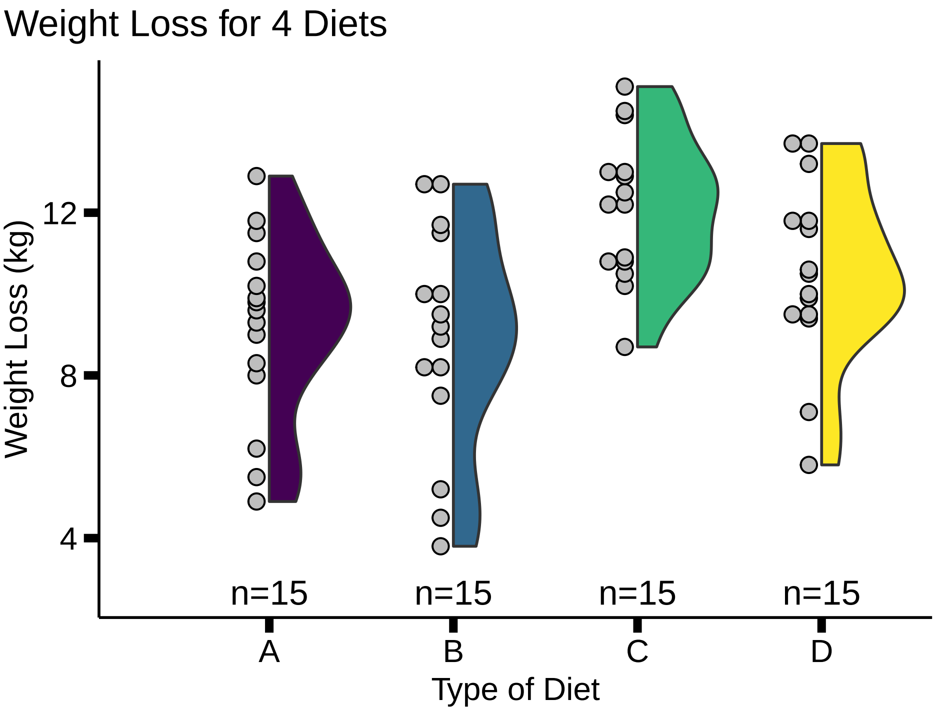
2.5 Explore the characteristics of distributions

The distributions can be explored visually with appropriate plots. Additionally, summary statistics and significance tests to check for normality (e.g., Shapiro-Wilk test) and for equality of variances (e.g., Levene's test) can be used.

2.5.1 Visualization of the distributions

We can visualize the distribution of `WeightLoss` for the four `Diet` groups:

```
dataDWL %>%
  ggplot(aes(x = Diet, y = WeightLoss, fill = Diet)) +
  geom_violindot(fill_dots = "grey", color_dots = "black", size_dots = 8) +
  stat_n_text(size = 4.5) +
  scale_fill_viridis_d() +
  labs( x = "Type of Diet", y = "Weight Loss (kg)",
       title = "Weight Loss for 4 Diets") +
  theme_pubr() +
  theme(plot.title.position = "plot",
        legend.position = "none",
        axis.ticks = element_line(size = 1.5,color="black"),
           axis.title = element_text(size = 12),
           axis.text = element_text(size = 12),
           axis.ticks.length=unit(0.2,"cm"))
```



The above figure shows that the data are close to symmetry and the assumption of a normal distribution is reasonable. Additionally, we can observe that the largest weight loss seems to have been achieved by the participants in `C` diet.

2.5.2 Summary statistics

The `WeightLoss` summary statistics for each diet group are:

```
DWL_summary <- dataDWL %>%
  group_by(Diet) %>%
  dplyr::summarize(
    n = n(),
    min = min(WeightLoss, na.rm = TRUE),
    q1 = quantile(WeightLoss, 0.25, na.rm = TRUE),
    median = quantile(WeightLoss, 0.5, na.rm = TRUE),
    q3 = quantile(WeightLoss, 0.75, na.rm = TRUE),
    max = max(WeightLoss, na.rm = TRUE),
    mean = mean(WeightLoss, na.rm = TRUE),
    sd = sd(WeightLoss, na.rm = TRUE),
    skewness = skewness(WeightLoss, na.rm = TRUE),
    kurtosis= kurtosis(WeightLoss, na.rm = TRUE)
  ) %>%
  ungroup()

DWL_summary
```

Diet	n	min	q1	median	q3	max
A	15	4.9	8.15	9.6	10.50	12.9
B	15	3.8	7.85	9.2	10.75	12.7
C	15	8.7	10.80	12.2	13.00	15.1
D	15	5.8	9.50	10.5	11.80	13.7

Diet	mean	sd	skewness	kurtosis
A	9.180000	2.295710	-0.4705790	-0.3020369
B	8.906667	2.781949	-0.4666645	-0.5153276
C	12.113333	1.793586	-0.0450739	-0.5301743
D	10.540000	2.233127	-0.4753756	0.2293477

The means are close to medians and the standard deviations are also similar. The skewness is approximately zero (symmetric distribution) and the (excess) kurtosis is close to zero (mesokurtic distribution) indicating normal distributions for all groups.

2.5.3 Shapiro-Wilk test for normality

The `Shapiro-Wilk` test for normality for each group is:

```
dataDWL %>%
  group_by(Diet) %>%
  shapiro_test(WeightLoss) %>%
  ungroup()
```

Diet	variable	statistic	p
A	WeightLoss	0.958	0.662
B	WeightLoss	0.941	0.390
C	WeightLoss	0.964	0.768
D	WeightLoss	0.944	0.435

The tests of normality suggest that the data for the `WeightLoss` in all groups are normally distributed (p > 0.05).

2.5.4 Levene's test for equality of variances

```
dataDWL %>%
      levene_test(WeightLoss ~ Diet)
```

df1	df2	statistic	p
3	56	0.6	0.617

Since the p-value = 0.617 >0.05, the null hypothesis that the variances of WeighLoss in four groups are equal can not be rejected.

2.6 Run the ANOVA test

Now, we will perform an one-way ANOVA (with equal variances: Fisher's classic ANOVA) to test the null hypothesis that the mean weight loss is the same for all the groups.

```
dataDWL %>%
  anova_test(WeightLoss ~ Diet, detailed = T)
```

Effect	SSn	SSd	DFn	DFd
Diet	97.33	296.987	3	56

F	p	p<.05	ges
6.118	0.001	*	0.247

F= 6.118 indicates the obtained F-statistic= (variation between sample means / variation within the samples). Note taht we are comparing to an F-distribution (F-test). The degrees of freedom in the numerator (DFn) and the denominator (DFd) are 3 and 56, respectively (numarator: variation between sample means; denominator: variation within the samples).

The p-value=0.001 is lower than 0.05. There is at least one diet with mean weight loss which is different from the others means.

From ANOVA table we can also calculate generalized effect size (ges). The `ges` is the proportion of variability explained by the factor `Diet` (SSn) to total variability of the dependent variable (SSn + SSd), so:

$$ges = 97.33/(97.33 + 296.987) = 97.33/394.317 = 0.247$$

A `ges` of 0.247 (24.7%) means that 24.7% of the change in the weight loss can be accounted for the diet conditions.

A summary table can also be presented:

Characteristic	**A**, N = 15[1]	**B**, N = 15[1]	**C**, N = 15[1]	**D**, N = 15[1]	**p-value**[2]
Weight Loss (kg)	9.2 (2.3)	8.9 (2.8)	12.1 (1.8)	10.5 (2.2)	0.001

[1] Mean (SD)

[2] One-way ANOVA

2.7 Post-hoc analysis (Tukey test)

A significant one-way ANOVA is generally followed up by Tukey post-hoc tests to perform multiple pairwise comparisons between groups:

```
# Pairwise comparisons
pwc_Tukey <- dataDWL %>%
  tukey_hsd(WeightLoss ~ Diet) %>%
  select(-null.value)

pwc_Tukey
```

term	group1	group2	estimate	conf.low	conf.high	p.adj	p.adj.signif
Diet	A	B	-0.2733333	-2.4999391	1.9532725	0.98800	ns
Diet	A	C	2.9333333	0.7067275	5.1599391	0.00513	**
Diet	A	D	1.3600000	-0.8666058	3.5866058	0.37700	ns
Diet	B	C	3.2066667	0.9800609	5.4332725	0.00190	**
Diet	B	D	1.6333333	-0.5932725	3.8599391	0.22200	ns
Diet	C	D	-1.5733333	-3.7999391	0.6532725	0.25200	ns

The output contains the following columns:

- estimate: estimate of the difference between means of the two groups
- conf.low, conf.high: the lower and the upper end point of the confidence interval at 95% (default)
- p.adj: p-value after adjustment for the multiple comparisons.

Pairwise comparisons were carried out using the method of Tukey and the adjusted p-values were calculated. The weight loss from diet C seems to be significantly larger than diet A (mean difference = 2.9 kg, 95%CI [0.71, 5.16], p=0.005 <0.05) and diet B (mean difference = 3.2 kg, 95%CI [0.98, 5.43], p=0.002 <0.05).

2.8 T-tests with Bonferroni Correction

Alternatively, we can perform pairwise comparisons using pairwise t-test with the assumption of equal variances (`pool.sd = TRUE`) and calculate the adjusted p-values using Bonferroni correction:

```
pwc_Bonferroni <- dataDWL %>%
  pairwise_t_test(
    WeightLoss ~ Diet, pool.sd = TRUE,
    p.adjust.method = "bonferroni"
    )
pwc_Bonferroni
```

.y.	group1	group2	n1	n2	statistic	df	p	p.adj	p.adj.signif
WeightLoss	A	B	15	15	0.2934996	27.02659	0.771000	1.000	ns
WeightLoss	A	C	15	15	-3.8996341	26.45175	0.000593	0.004	**
WeightLoss	A	D	15	15	-1.6446418	27.97864	0.111000	0.666	ns
WeightLoss	B	C	15	15	-3.7520590	23.92403	0.000987	0.006	**
WeightLoss	B	D	15	15	-1.7732618	26.74879	0.088000	0.526	ns
WeightLoss	C	D	15	15	2.1274462	26.75471	0.043000	0.257	ns

3 ANOVA with unequal variances

3.1 Welch ANOVA

If the variance is different between the groups (unequal variances) then the degrees of freedom associated with the ANOVA test are calculated differently (Welch ANOVA).

```
# Welch one-way ANOVA test (not assuming equal variance)


dataDWL %>%
  welch_anova_test(WeightLoss ~ Diet)
```

.y.	n	statistic	DFn	DFd	p	method
WeightLoss	60	7.02	3	30.77044	0.000989	Welch ANOVA

In this case, the `Games-Howell post hoc test` (or pairwise t-tests with no assumption of equal variances with Bonferroni correction) can be used to compare all possible combinations of group differences.

3.2 Games-Howell post hoc test

```
# Pairwise comparisons (Games-Howell)
pwc_GH <- dataDWL %>%
  games_howell_test(WeightLoss ~ Diet)


pwc_GH
```

.y.	group1	group2	estimate	conf.low	conf.high	p.adj	p.adj.signif
WeightLoss	A	B	-0.2733333	-2.8217079	2.2750412	0.991	ns
WeightLoss	A	C	2.9333333	0.8721336	4.9945330	0.003	**
WeightLoss	A	D	1.3600000	-0.8978769	3.6178769	0.371	ns
WeightLoss	B	C	3.2066667	0.8485025	5.5648309	0.005	**
WeightLoss	B	D	1.6333333	-0.8888050	4.1554717	0.308	ns
WeightLoss	C	D	-1.5733333	-3.5983233	0.4516566	0.170	ns

4 Kruskal-Wallis test

The Kruskal-Wallis test is a rank-based nonparametric alternative to the one-way ANOVA and an extension of the Wilcoxon-Mann-Whitney test to allow the comparison of more than two independent groups. It's usually recommended when the assumptions of one-way ANOVA test are not met (non-normal distributions) or with small samples.

4.1 Research question

We wish to compare the VO2max in three different sports (runners, rowers, and triathletes).

4.2 H0 and H1 Hypotheses

- H_0: the distribution of VO2max is the same in all groups (the medians of VO2max in the three sports are the same)
- H_1: there is at least one group with VO2max distribution different from the others (there is at least one sport with median VO2max different from the others)

Note The Kruskal-Wallis test should be regarded as a test of dominance between distributions comparing the mean ranks. The null hypothesis is that the observations from one group do not tend to have a higher or lower ranking than observations from the other groups. This test **does not** test the medians of the data as is commonly thought, it tests the **whole distribution**. However, if the distributions of the two groups have **similar shapes**, the Kruskal-Wallis test can be used to determine whether there are differences in the medians in the two groups. In practice, we use the medians to present the results.

4.3 Preraring the data

We import the data:

```
library(readxl)
dataVO2 <- read_excel(here("data", "dataVO2.xlsx"), col_names=TRUE)
dataVO2
```

VO2max Data (first and last 5 rows)

sport	VO2max
runners	73.8
runners	79.9
runners	75.5
runners	72.5
runners	82.2
NA	...
triathletes	63.2
triathletes	65.8
triathletes	63.4
triathletes	65
triathletes	67

We inspect the data:

```
glimpse(dataVO2)
```

```
## Rows: 30
## Columns: 2
## $ sport  <chr> "runners", "runners", "runners", "runners", "runners", "ru~
## $ VO2max <dbl> 73.8, 79.9, 75.5, 72.5, 82.2, 78.3, 77.9, 76.5, 72.3, 80.2~
```

The dataset dataV02 has 30 participants and includes two variables. The numeric V02max variable and the sport variable (with levels "roweres", "runners", and "triathletes") which should be converted from character to a factor variable using the factor():

```
dataV02 <- dataV02 %>%
  mutate(sport = factor(sport))
glimpse(dataV02)
```

```
## Rows: 30
## Columns: 2
## $ sport  <fct> runners, runners, runners, runners, runners, runners, runn~
## $ V02max <dbl> 73.8, 79.9, 75.5, 72.5, 82.2, 78.3, 77.9, 76.5, 72.3, 80.2~
```

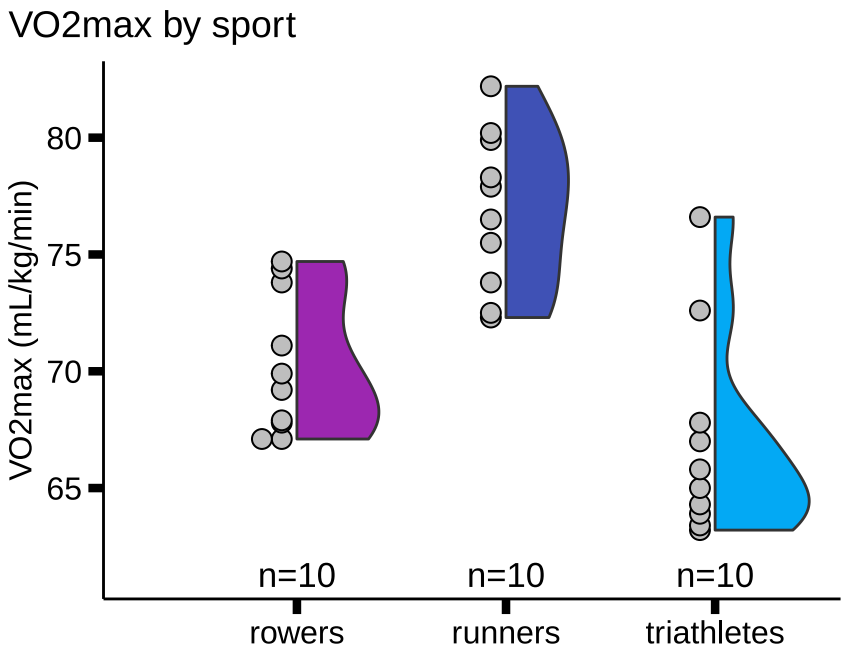
4.4 Explore the characteristics of distributions

The distributions of the three groups can be checked with **dot plots and histograms** and summary statistics.

4.4.1 Visualization of the distributions

We can visualize the distribution of VO2max for the three groups:

```
dataV02 %>%
  ggplot(aes(x = sport, y = V02max, fill = sport)) +
  geom_violindot(fill_dots = "grey", color_dots = "black", size_dots = 17) +
   scale_fill_material_d(palette = "ice") +
  labs(y = "V02max (mL/kg/min)",
       title = "VO2max by sport") +
  stat_n_text(size = 4.5) +
  theme_pubr() +
  theme(plot.title.position = "plot",
        legend.position = "none",
        axis.ticks = element_line(size = 1.5, color="black"),
           axis.title.x = element_blank(),
           axis.title = element_text(size = 12),
           axis.text = element_text(size = 12),
           axis.ticks.length=unit(0.2,"cm"))
```



The above figure shows that the data in triathletes group have some outliers. Additionally, we can observe that the runners group seems to have the largest VO2max.

4.4.2 Summary statistics

Summary statistics can also be inspected in each sport:

```
VO2_summary <- dataVO2 %>%
  group_by(sport) %>%
  dplyr::summarize(
    n = n(),
    min = min(VO2max, na.rm = TRUE),
    q1 = quantile(VO2max, 0.25, na.rm = TRUE),
    median = quantile(VO2max, 0.5, na.rm = TRUE),
    q3 = quantile(VO2max, 0.75, na.rm = TRUE),
    max = max(VO2max, na.rm = TRUE),
    mean = mean(VO2max, na.rm = TRUE),
    sd = sd(VO2max, na.rm = TRUE),
    skewness = skewness(VO2max, na.rm = TRUE),
    kurtosis= kurtosis(VO2max, na.rm = TRUE)
  ) %>%
  ungroup()

VO2_summary
```

sport	n	min	q1	median	q3	max
rowers	10	67.1	67.825	69.55	73.125	74.7
runners	10	72.3	74.225	77.20	79.500	82.2
triathletes	10	63.2	64.000	65.40	67.600	76.6

sport	mean	sd	skewness	kurtosis
rowers	70.30	3.035347	0.5021634	-1.533375
runners	76.91	3.386066	-0.0095043	-1.164338
triathletes	66.96	4.395503	1.5062665	1.602985

The sample size is relative small. Moreover, the skewness and the kurtosis for the triathletes are higher than zero indicating non-normal leptokurtic distribution.

4.4.3 Shapiro-Wilk test for normality

Additionally, we can check the normality applying the `Shapiro-Wilk` test:

```
dataVO2 %>%
  group_by(sport) %>%
  shapiro_test(VO2max) %>%
  ungroup()
```

sport	variable	statistic	p
rowers	VO2max	0.865	0.087
runners	VO2max	0.954	0.712
triathletes	VO2max	0.816	0.023

We can see that the data for the triathletes is not normally distributed (p=0.023 <0.05) according to the Shapiro-Wilk test.

By considering all of the information together (small samples, graphs, normality test) the overall decision is against of normality.

4.5 Run the Kruskal-Wallis test

Now, we will perform a Kruskal-Wallis test to compare the VO2max in three sports.

```
dataVO2 %>%
  kruskal_test(VO2max ~ sport)
```

.y.	n	statistic	df	p	method
VO2max	30	16.35091	2	0.000281	Kruskal-Wallis

The p-value (<0.001) is lower than 0.05. There is at least one sport in which the VO2max is different from the others.

A summary table can also be presented:

Characteristic	**rowers**, N = 10[1]	**runners**, N = 10[1]	**triathletes**, N = 10[1]	**p-value**[2]
VO2max (mL/kg/min)	69.6 (67.8, 73.1)	77.2 (74.2, 79.5)	65.4 (64.0, 67.6)	<0.001

[1] Median (IQR)

[2] Kruskal-Wallis rank sum test

4.6 Post-hoc analysis (Dunn's approach)

A significant WMW is generally followed up by Dunn post-hoc tests to perform multiple pairwise comparisons between groups:

```
# Pairwise comparisons
pwc_Dunn <- dataVO2 %>%
  dunn_test(VO2max ~ sport, p.adjust.method = "bonferroni")


pwc_Dunn
```

.y.	group1	group2	n1	n2	statistic	p	p.adj	p.adj.signif
VO2max	rowers	runners	10	10	2.426512	0.0152447	0.0457342	*
VO2max	rowers	triathletes	10	10	-1.588032	0.1122792	0.3368376	ns
VO2max	runners	triathletes	10	10	-4.014544	0.0000596	0.0001787	***

Dunn's pairwise comparisons were carried out using the method of Bonferroni and adjusting the p-values were calculated. The runners' VO2max (median= 77.2, IQR=[74.2, 79.5] mL/kg/min) seems to differ significantly (larger based on the medians) from rowers (69.6 [67.8, 73.1] mL/kg/min, p=0.046 <0.05) and triathletes (65.4 [64.0, 67.6] mL/kg/min, p <0.001).

4.7 Pairwise comparisons using WMW's test with Bonferroni correction

Alternatively, we can perform pairwise comparisons using pairwise WMW's test and calculate the adjusted p-values using Bonferroni correction:

```
# Pairwise comparisons
pwc_BW <- dataVO2 %>%
  pairwise_wilcox_test(VO2max ~ sport, p.adjust.method = "bonferroni")


pwc_BW
```

.y.	group1	group2	n1	n2	statistic	p	p.adj	p.adj.signif
VO2max	rowers	runners	10	10	8.5	0.002000	0.006	**
VO2max	rowers	triathletes	10	10	80.5	0.023000	0.070	ns
VO2max	runners	triathletes	10	10	93.0	0.000487	0.001	**