One-way ANOVA: Testing the homogeneity of more than 2 means

Josep L. Carrasco
Bioestadística. Departament de Fonaments Clínics
Universitat de Barcelona

Introduction

The goal of this lesson is to study homogeneity of means from k independent samples, where $k > 2$. Thus, our objective is to test the following null hypothesis:

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

against the alternative hypothesis that some means are different.

We could test this hypothesis by comparing each possible pair of means using the methods we saw in previous lessons. However, this approach has a serious drawback: the multiple comparisons problem. This occurs when one considers a set of statistical inferences simultaneously. If we compare one of the possible pairs of means, the probability of making a type I error equal to α. If we make multiple comparisons to compare them all, the global probability of type I error is (assuming that the tests are independent)

$$\alpha^* = 1 - (1 - \alpha)^c ,$$

where c is the number of comparisons performed, that is,

$$c = \frac{k\,(k-1)}{2}.$$

For example, if we want to compare 5 means, the number of comparisons to make is $c = \frac{5 \cdot 4}{2} = 10$. Each comparison is solved with a significance level $\alpha = 0.05$. If the samples are independent, the probability of not rejecting all the null hypotheses given that these are true (global level of confidence) is $0.95^{10} \simeq 0.60$. Thus, the global significance level is $1-(1-0.05)^{10} = 0.40$. Hence, the probability of incorrectly reject some of the null hypotheses is 0.40, which is considerably higher than the desired 0.05.

Thus, we need some methodology that allows us to evaluate simultaneously all means. This procedure is known as **an**alysis **o**f **va**riance or ANOVA. This is a generalization of the t test for independent data and allows us to work with any number of means.

Analysis of variance

Example data

In this study, weight at birth was registered in a sample of 656 newborns. We want to know if there are differences in weight at birth depending on whether this was the first labour of the mother, whether this was the second, third or fourth labour or whether this was the fifth or greater labour. Hence, our outcome of interest is a continous variable (weight at birth) and the factor identifying the groups to compare is the labour number, which has three categories.

First of all, we import the dataset to R:

```
weight <- read.table("pesnen.txt", header=T, sep="\t")
head(weight)
```

```
  paridad pes.nen
1       7     1.3
2       2     1.4
3       2     1.5
4       7     1.8
5       1     1.8
6       1     1.8
```

This data set contains the following variables:

- paridad: number of labour of the mother
- pes.nen: weight at birth of the newborn, in kg

We need to create a new variable that identifies the number of labour group:

```
max(weight$paridad)
```

```
[1] 11
```

```
weight$group <- cut(weight$paridad, breaks=c(0, 1, 4, 11),
                    labels=c("1", "2-4", "5 or more"))
table(weight$group)
```

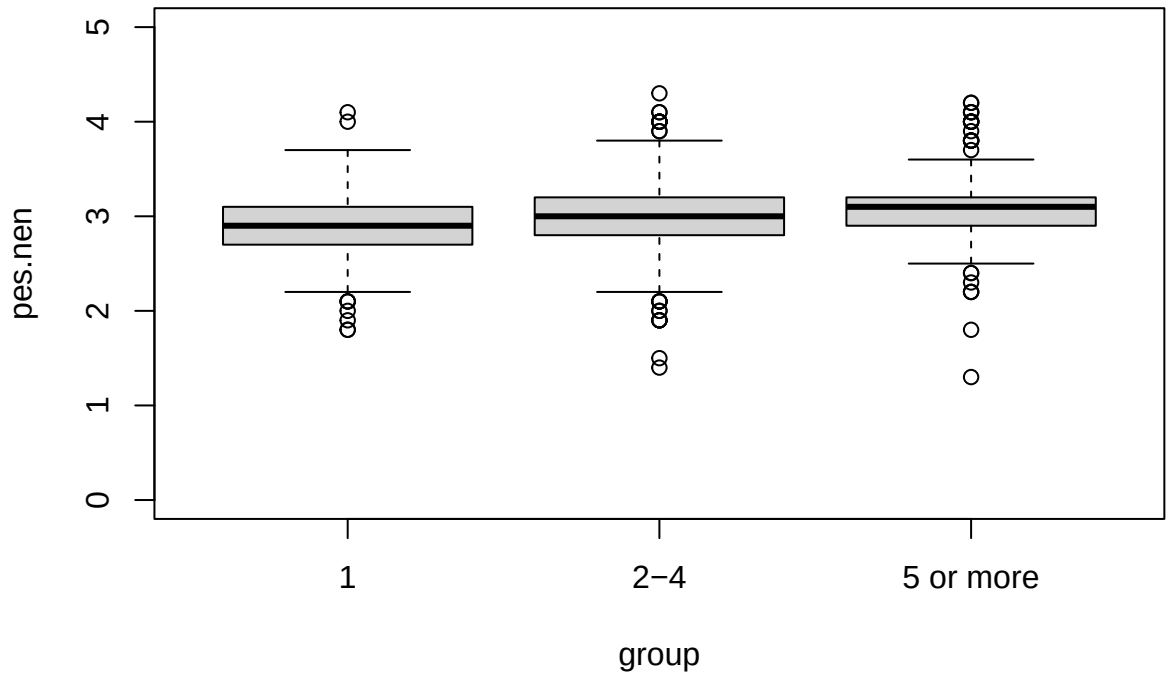
```

        1       2-4 5 or more
      105       340       211
```

Let us see some descriptive statistics about the birth weight according to the number of labour group:

```
by(weight$pes.nen, weight$group, summary)
```

```
weight$group: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.800   2.700   2.900   2.904   3.100   4.100
------------------------------------------------------------
weight$group: 2-4
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.400   2.800   3.000   3.016   3.200   4.300
------------------------------------------------------------
weight$group: 5 or more
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.300   2.900   3.100   3.082   3.200   4.200
```

```
boxplot(pes.nen~group, weight, ylim=c(0, 5))
```



Variance decomposition

ANOVA is based on decomposing the total variability of the data in two components:

- Between-groups variability: variability between each group's means with respect to the global mean (mean of the entire dataset).
- Within-groups variability: Variable of each individual with respect to the mean of its group.

High between-groups variability will indicate that the means are different among the groups and will contribute to reject the null hypothesis (equality of means among grups).

Variability of the data is expressed in terms of sum of squares. The total sum of squares (SS_T) is

$$SS_T = \sum_{i=1}^{k} \sum_{k=1}^{n} \left(y_{ij} - \overline{y}_{**}\right)^2,$$

where $\overline{y}_{**}$ indicates the global mean, y_{ij} is the outcome of individual j from group i, with $i = 1, \ldots, k$ (k groups) and $j = 1, \ldots, n$. Notice that we assume that the number of individuals is the same in all groups. This is known as *balanced* design. The ANOVA method is also valid for unbalanced designs (the number of subjects is different in each group) with minimal modifications, but in terms of statistical power and applicability conditions, balanced designs are preferable.

The total sum of squares can be decomposed in the sum of squares between groups (SS_B) and the sum of squares within-groups (SS_W).

$$\begin{aligned} SS_T &= SS_B + SS_W \\ &= \sum_{i=1}^{k} \left(\overline{y}_{i*} - \overline{y}_{**}\right)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n} \left(y_{ij} - \overline{y}_{i*}\right)^2 \end{aligned}$$

where $\overline{y}_{i*}$ represents the mean of the ith group.

Hypothesis test

Both between- and within-groups variability are associated to a certain number of degrees of freedom, which are respectively $k - 1$ and $n - k$, with a total of $n - 1$ degrees of freedom. If we divide the sums of squares by their respective degrees of freedom, we obtain the mean sums of squares (MS_B and MS_W):

$$MS_B = \frac{SS_B}{k-1}, \qquad MS_W = \frac{SS_W}{n-k}.$$

We can also compute the mean sum of squares as

$$MS_T = \frac{SS_T}{n-1} = \frac{\sum\limits_{i=1}^{k} \sum\limits_{k=1}^{n} \left(y_{ij} - \overline{y}_{**}\right)^2}{n-1},$$

which is the sample variance.

The expectation of the mean sum of squares is key to ANOVA. It can be shown that

$$E\left[MS_W\right] = \sigma^2,$$

$$E\left[MS_B\right] = \sigma^2 + n\sum_{i=1}^{k}\left(\mu_i - \mu\right)^2$$

where σ^2 represents the within-groups variance, μ_i is the population mean of the ith group and μ is the global mean.

If the null hypothesis of equal means among groups is true ($\mu_1 = \ldots = \mu_k = \mu$), then $\sum_{i=1}^{k}\left(\mu_i - \mu\right)^2 = 0$ and $E\left[MS_B\right] = E\left[MS_W\right]$. Thus, the ratio

$$F = \frac{MS_B}{MS_W}$$

can be used to evaluate the likelihood of the null hypothesis. Actually, F is the statistic that ANOVA uses to test the null hypothesis of equal means. If the null hypothesis is true, the probability distribution model of F is an F-distribution with $k-1$ and $k\left(n-1\right)$ degrees of freedom.

Let us execute the analysis with our example dataset:

```
model <- aov(pes.nen~group, weight)
summary(model)
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
group         2   2.22  1.1109   6.278 0.00199 **
Residuals   653 115.54  0.1769
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result of executing the function `aov` is known as ANOVA table. This table shows, for each source of variability, the degrees of freedom (`Df`), the sums of squares (`Sum Sq`), the mean sums of squares (`Mean Sq`), the statistic (`F value`) and the p-value (`Pr(>F)`). If we set $\alpha = 5\%$, since the p-value is lower than 5%, we reject the null hypothesis of equal means.

Applicability conditions

If we want the ANOVA test to be valid, the following applicability conditions must hold:

1) The data in each group must come from independent samples and must follow a Normal distribution. This condition can be relaxed if the sample size is greater than 30 in each group.

The normality condition can be evaluated by looking at the model residuals, that is, the difference between each observation and the mean of its group: $y_{ij} - \overline{y}_{i*}$.

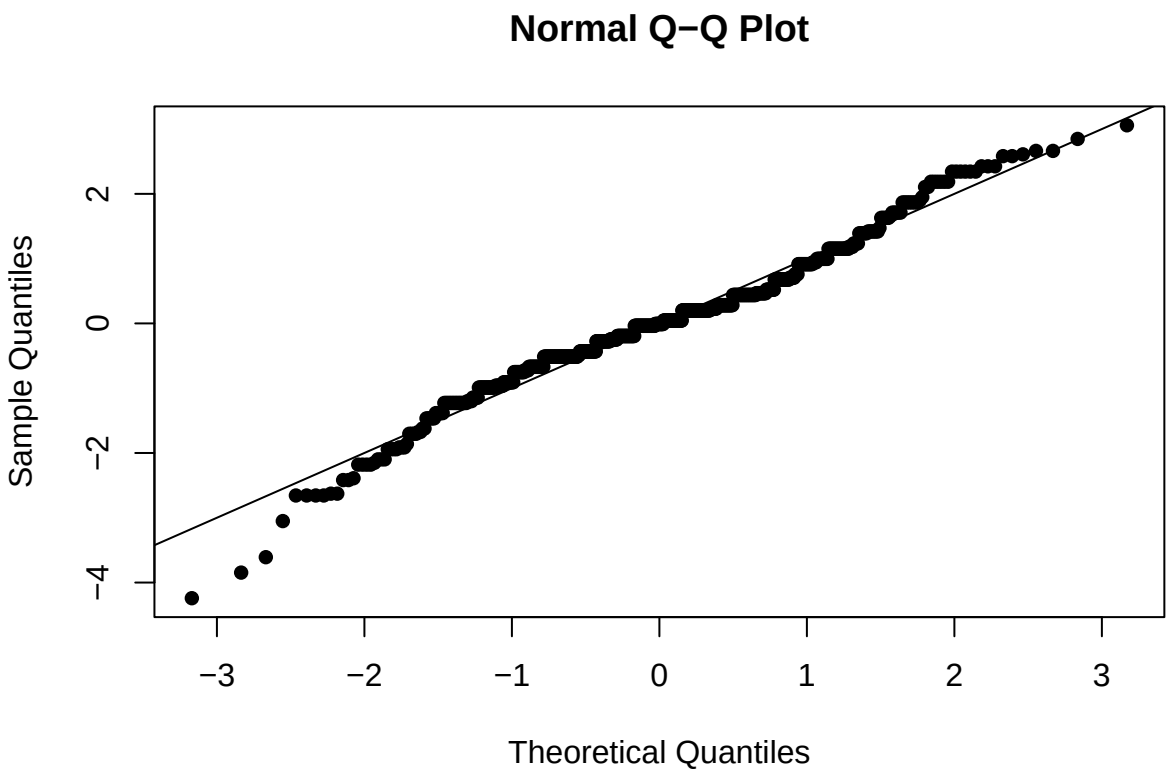
Let us apply the Shapiro-Wilks test to our residuals:

```
res <- residuals(model)
shapiro.test(res)
```

```
    Shapiro-Wilk normality test

data:  res
W = 0.97637, p-value = 8.5e-09
```

This test gives us a significant result (using a significance level of 5%). But is this departure from normality so serious? Let us make a Q-Q plot to check it:

```
qqnorm((res-mean(res))/sd(res),pch=16)
abline(0, 1)
```



The Q-Q plot shows that the distribution of residuals is fairly Normal except for a few observations at the queues, which causes the Shapiro-Wilks test to be significant. Thus, we

conclude that although the residuals cannot be said to follow a Normal distribution, they are not so far from it. Moreover, the sample size is greater than 30 in the three groups, so we assume that this first applicability condition holds.

2) Homogeneity of variances among the groups. To check this condition we need to test the following null hypothesis:

$$H_0 : \sigma_1 = ... = \sigma_k = \sigma,$$

against the alternative hypothesis that some variances are different. To test these hypothesis we have two options: **Bartlett test** and **Fligner-Killeen test**.

The Bartlett test is appropriate when the outcome of interest follows a Normal distribution. We have seen that the assumption of normality have been met. Therefore, let us use this test to chech the equality of variances.

```
bartlett.test(pes.nen~group, weight)
```

```
    Bartlett test of homogeneity of variances

data:  pes.nen by group
Bartlett's K-squared = 0.92901, df = 2, p-value = 0.6284
```

The p-value is greater than 5%, so we do not reject the null hypothesis of equal variances.

A posteriori multiple comparisons

When the null hypothesis of equal means is rejected, we conclude that some means are different. However, ANOVA does not identify which groups are different. The methodology to discover which groups are different are called **multiple comparisons**, **post-hoc comparisons** or **a posteriori** comparisons. The main challenge of this problem is to look for differences across groups while maintaining the global significance level α, that is, we need to avoid the multiple comparisons problem we saw before. In statistical literature we can find several methods to achieve this. Here we are going to see two approaches: the Bonferroni method and the Tukey's Honestly Significant Difference (HSD).

- **Bonferroni correction:** This method has become popular due to its simplicity. The idea is to correct the type I error probability of every comparison performed. Thus, we test the global hypothesis of equal means using the ANOVA procedure. If this test is significant, we compare every possible pair of means using a t test, but we use a

corrected value of α. Bonferroni proposed to divide α by the number of comparisons performed. Recall that, in our example,

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

so we need to perform three a posteriori comparisons:

$$\mu_1 = \mu_2$$
$$\mu_1 = \mu_3$$
$$\mu_2 = \mu_3$$

To use the Bonferroni method, we need to solve each hypothesis test with the following significance level:

$$\alpha^* = \frac{\alpha}{3} = \frac{0.05}{3} = 0.0167.$$

Let us do it with our example dataset:

```
x1 <- subset(weight, group=="1")$pes.nen # sample 1
x2 <- subset(weight, group=="2-4")$pes.nen # sample 2
x3 <- subset(weight, group=="5 or more")$pes.nen # sample 3
t.test(x1, x2)
```

```
    Welch Two Sample t-test

data:  x1 and x2
t = -2.3411, df = 173.44, p-value = 0.02036
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.20601659 -0.01754083
sample estimates:
mean of x mean of y
 2.903810  3.015588
```

```
t.test(x2, x3)
```

```
    Welch Two Sample t-test

data:  x2 and x3
t = -1.8185, df = 464.88, p-value = 0.06963
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
```

```
 -0.137171189  0.005314485
sample estimates:
mean of x mean of y
 3.015588  3.081517
```

```
t.test(x1, x3)
```

```
    Welch Two Sample t-test

data:  x1 and x3
t = -3.544, df = 197.79, p-value = 0.0004917
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.27658984 -0.07882429
sample estimates:
mean of x mean of y
 2.903810  3.081517
```

So the only significant difference we find in this example is between μ_1 and μ_3. The remaining two p-values are low but not lower than 0.0167.

The main disadvantage of the Bonferroni correction is that it can be too conservative, that is, with low power to detect significant differences, specially if there are a large number of tests. This could lead to situations in which we find global differences with the ANOVA test but then the Bonferroni method fails to find any significant difference.

- **Tukey's HSD:** This procedure is based on ordering the means and comparing them. First, we compare the largest mean with the lowest one. If the difference is statistically significant, we continue with the second lowest mean, and we repeat this step with every mean. Then, the procedure is repeated with the second largest mean. The algorithm ends when all means have been compared or when we find a non-significant difference, since at this point we consider that the remaining differences are non-significant.

In R:

```
TukeyHSD(model)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = pes.nen ~ group, data = weight)
```

```
$group
                     diff           lwr       upr     p adj
2-4-1          0.11177871  0.001457802 0.2220996 0.0462245
5 or more-1    0.17770706  0.059696849 0.2957173 0.0012568
5 or more-2-4 0.06592835 -0.020669485 0.1525262 0.1742560
```

This method results in significant differences between μ_1 and μ_2 and between μ_1 and μ_3, but not between μ_2 and μ_3. Notice that with this method we are able to find a difference that was not detected with the Bonferroni method (the Tukey method is less conservative, thus it has more power than the Bonferroni method).

As a conclusion in the context of our example, we can say that the effect of the number of previous labours on the birth weight is statisically significant. These differences in mean birth weights are found between mothers in their first labour and the rest of mothers.

Alternative procedures when the applicability conditions do not hold

In a study about fertility, the reseachers wanted to evaluate the relationship between the follicle stimulating hormone (FSH) and the quality of sperm. With this aim, the ratio between FSH and the inhibin segregated was compared in three groups of male individuals. These three groups were determined from the count of spermatozoons in the ejaculation. Thus, these subjects were classified in oligoasthenozoospermia, moderate asthenozoospermia and severe asthenozoospermia.

Let us import the data to R:

```
fsh <- read.table("fshrat.txt", header=T, sep="\t")
```

The first 10 individuals:

```
fsh[1:10, ]
```

```
        rati       grup
1  4.3854167 A.moderats
2  7.8275862 A.moderats
3  5.1694915 A.moderats
4  6.6153846 A.moderats
5  2.6170213 A.moderats
6  2.5357143 A.moderats
7  1.7727273 A.moderats
8  1.3846154 A.moderats
9  0.7607143 A.moderats
10 0.2666667 A.moderats
```

Number of individuals per group:

```
table(fsh$grup)
```

```
A.moderats   A.severs      Oligo
        16         14         44
```

Summary descriptives per group:

```
by(fsh$rati, fsh$grup, summary)
```

```
fsh$grup: A.moderats
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2667  0.9684  1.7073  2.5531  3.0591  7.8276
------------------------------------------------------------
fsh$grup: A.severs
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.176   1.708   2.445   2.988   4.387   5.591
------------------------------------------------------------
fsh$grup: Oligo
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6316  2.0701  3.0851  3.6124  4.2089 20.0000
```

The hypothesis to solve is: equality of ratio means according to the quality of sperm group. Thus,

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

First of all, let us evaluate the applicability conditions of the ANOVA test.

- Normality of the ratio:

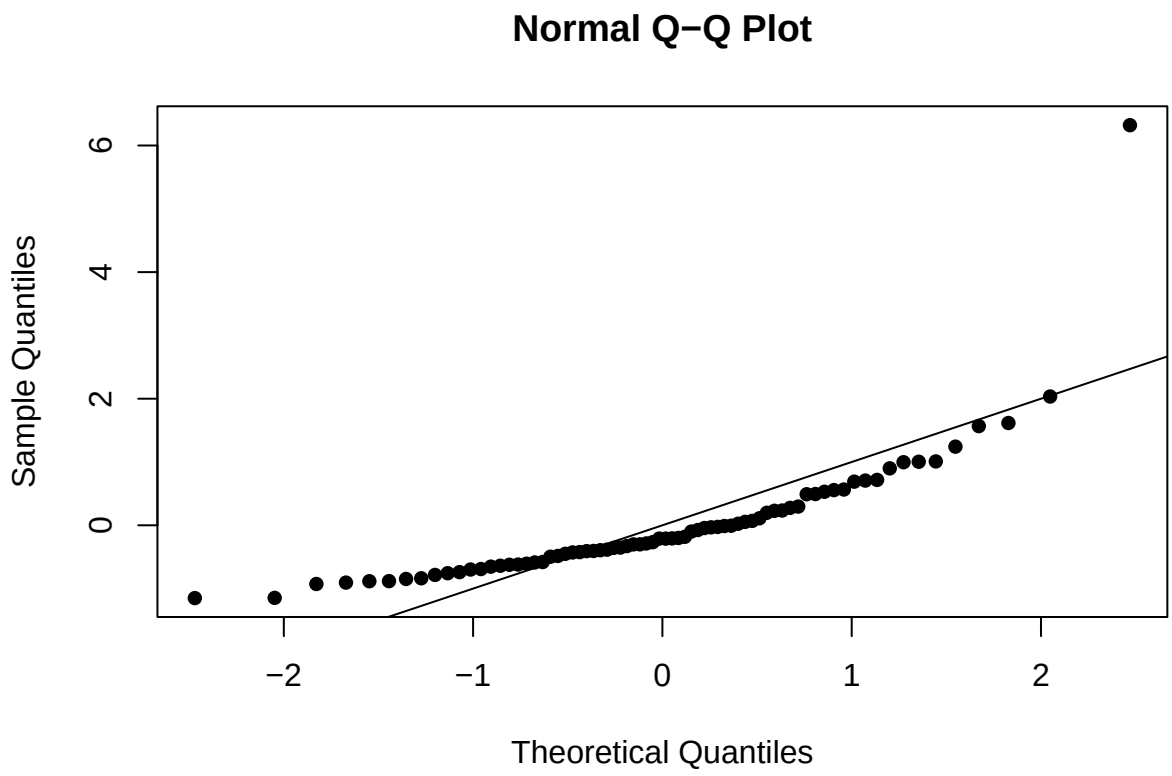
```
model <- aov(rati~grup, data=fsh)
summary(model)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
grup         2   14.5   7.249   1.049  0.356
Residuals   71  490.7   6.911
```

```
res <- residuals(model)
shapiro.test(res)
```

```
    Shapiro-Wilk normality test

data:  res
W = 0.69887, p-value = 4.97e-11
```

```
qqnorm((res-mean(res))/sd(res),pch=16)
abline(0, 1)
```



Both the Shapiro-Wilks test and the QQ-plot show that the assumption of normality is not acceptable. Moreover, the sample size is small in two of the groups.

Thus, the **Fligner-Killeen test** will be the choice to check the equality of variance .

- Homogeneity of variances:

```
fligner.test(rati~grup, data=fsh)
```

```
    Fligner-Killeen test of homogeneity of variances

data:  rati by grup
Fligner-Killeen:med chi-squared = 0.21042, df = 2, p-value = 0.9001
```

The null hypothesis of equality of variances is not rejected.

However, the applicability conditions for ANOVA test do not hold in this example beucause the normality assumption. So we propose two alternative approaches.

Permutation test

The permutation test is based on randomly exchanging the data. The idea behind this is the following: if all the groups come from populations with the same mean, the group assignation is trivial. In each permutation we calculate the contrast statistic, such that when we repeat this process a high number of times we will be generating the distribution of the statistic under the null hypothesis.

Let us calculate the contrast statistic (in this case, the F statistic) in our sample:

```
model <- aov(rati~grup, data=fsh)
summary(model)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
grup         2   14.5   7.249   1.049  0.356
Residuals   71  490.7   6.911
```

```
theta.star <- summary(model)[[1]][1, 4]
theta.star
```

```
[1] 1.048954
```

Then we create a function to 1) generate a permutation, 2) calculate the F statistic and 3) return the result.

```
perm.anova <- function(){
  # new dataset, with permutation
  newdat <- data.frame(sample(fsh$rati), fsh$grup)
  names(newdat) <- c("rati", "grup")
  model <- aov(rati~grup, data=newdat) # ANOVA
  theta <- summary(model)[[1]][1, 4] # store F
  theta
}
```

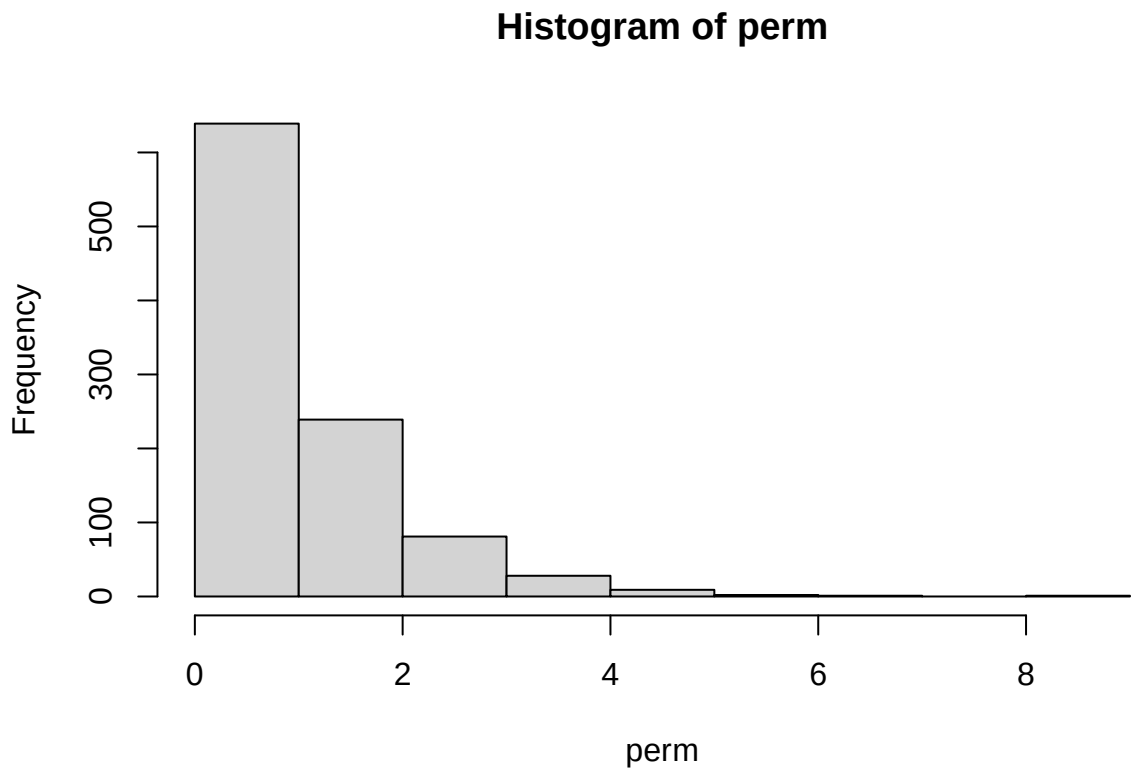
Let us generate 1000 permutations and see the first 20:

```
perm <- replicate(1000, perm.anova())
perm[1:20]
```

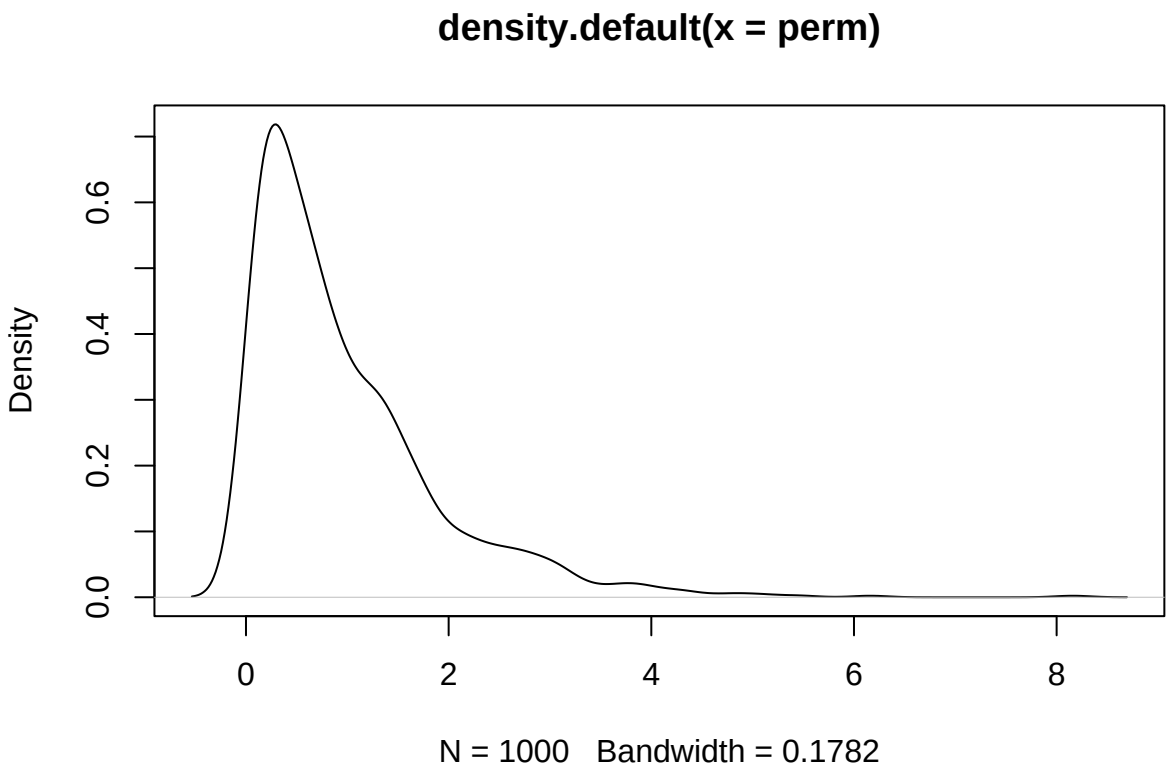
```
 [1] 1.32292167 0.17128344 1.22832286 0.08993485 1.84173055 0.41391877
 [7] 0.74486334 0.93050221 0.24778084 0.39994904 2.14133286 1.26809226
[13] 1.67402798 0.92424188 0.63831031 3.17274145 1.41690404 0.13514301
[19] 1.43431102 0.03732983
```

The distribution of the statistic under the null hypothesis is

```
hist(perm)
```




```
plot(density(perm))
```



and finally we calculate the p-value as the proportion of values greater than the statistic in our sample:

```
(sum(perm>theta.star)+1)/(1000+1)
```

```
[1] 0.3506494
```

Our p-value is greater than 5%, so we do no reject the null hypothesis of equality of means.

It is worthy to note that if the null hypothesis would have been rejected, the post-hoc comparisons should have been done as permutation tests using t-test statistic.

Kruskal-Wallis test

The Kruskal-Wallis test is a generalization of the Mann-Whitney test for the case of more than 2 groups. The null hypothesis is equality of medians among the populations under comparison. The `R` syntax for this test is:

```
kruskal.test(rati~grup, data=fsh)
```

```
    Kruskal-Wallis rank sum test

data:  rati by grup
Kruskal-Wallis chi-squared = 4.1457, df = 2, p-value = 0.1258
```

This test also gives us a p-value greater than 5%, so we do no reject the null hypothesis of equality. In the case of a significant p-value, the post-hoc comparisons could be done with Mann-Whitney tests for every pair of groups and then applying the Bonferroni correction.

Sample size

The function *pwr.anova.test(k,f,power)* from *pwr* package can be used to compute the sample size necessary to reject the null hypothesis with a certain power.

The arguments of the function are:

- *k*. Number of groups.
- *f*. Effect size. It is computed as
$$f = \frac{SS_B}{SS_T}$$
it is assumed that $f = 0.1$ stands for a small effect, whereas 0.25 and 0.4 involve medium and large effects respectively.
- *power*. Power of test

For example, what is the sample size necessary to detect a small effect with 3 groups and power of 80%?

```
library(pwr)
pwr.anova.test(k=3,f=0.1,power=0.8)
```

```
     Balanced one-way analysis of variance power calculation

              k = 3
              n = 322.157
              f = 0.1
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

The result is 323 subjects **in each group**.