Example: Analysis of Variance (ANOVA) and alternative methods

XDASI Fall 2021

November 15, 2021

ANOVA Example

The following example is from these Khan Academy videos:

- Calculating SST
- Calculating SSW and SSB
- Hypothesis Testing

We previously discussed how we can use the t-test to determine if two sample distributions come from populations with the same mean (in which case, assuming equal variances, we can say that they come from the same population).

In many cases, we will have ***multiple*** sample groups and we will want to ask a similar question:

Are the means of the different samples the same ?

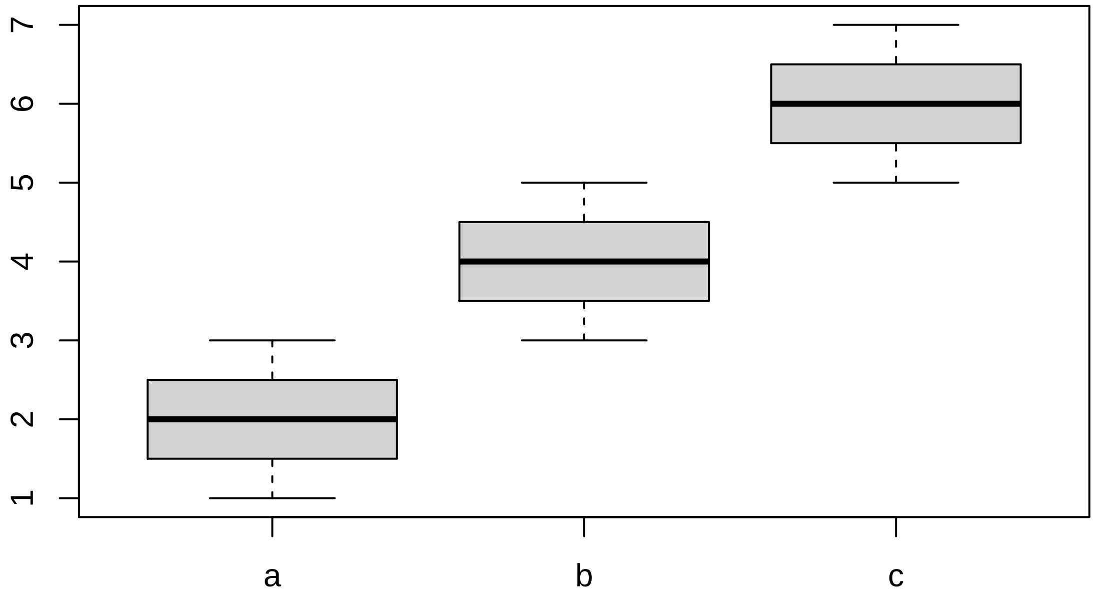
To answer this question we will look at a very simple case with three conditions – a, b, and c – and ask if their means are significantly different.

```
# measurements for three conditions
a=c(3,2,1)
b=c(5,3,4)
c=c(5,6,7)

anova_mat = cbind(a,b,c)  # combine the data into a 3x3 matrix
anova_mat                 # take a look at the matrix
```

```
##      a b c
## [1,] 3 5 5
## [2,] 2 3 6
## [3,] 1 4 7
```

```
boxplot(as.data.frame(anova_mat)) # plot it as a data frame
```



Looking at the boxplots above, it is clear to see that their means are indeed different. So the question we want to ask is whether the differences are ***significant***.

Sums of squares

Instead of looking at the difference between the sample means, as we did with t-test, we will compare variances. There are three different variances that we can calculate:

- ***SST*** (Total Sum of Squares) = variation of all the points to the overall mean.
- ***SSW*** (Within Group Sum of Squares) = variation of the data within each group.
- ***SSB*** (Between Group Sum of Squares) = variation of the group mean to the overall mean.

We also need the ***degrees of freedom***. Given that you know the average, how many values you need to know? It's simply one less than the number of items being considered for each comparison, because using the mean you can always calculate the last value.

To calculate **SST**, we simply take the difference of all the values from the overall mean, square them, and then take the sum.

$$SST = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \bar{X})^2$$

```
# overall mean of the data
anova_mat_mean = mean(anova_mat)

# total variation = sum of squared deviations
#                   of each data point from the overall mean
SST = sum((anova_mat - anova_mat_mean)**2)
SST
```

```
## [1] 30
```

Since this is a sample of the entire population, our degrees of freedom equal the total number of values minus one.

```
# total degrees of freedom = (# of data points) - 1
SST_df = length(anova_mat)-1
SST_df
```

```
## [1] 8
```

SSW (Within Group Sum of Squares) = variation of the data within each group. Here we calculate the variation of each point relative to the mean of its own group and simply add up the squared differences across all the groups:

$$SSW = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \bar{X}_j)^2$$

where n is the number of measurements in each group, and m is the number of groups

```
anova_mat_col_mean = colMeans(anova_mat)
anova_mat_col_mean
```

```
## a b c
## 2 4 6
```

```
SSW=0
for ( i in 1:nrow(anova_mat)) {
  SSW = SSW + sum((anova_mat[i,]-anova_mat_col_mean)**2)
}
SSW
```

```
## [1] 6
```

When calculating the degree of freedom, remember that we calculated the sum of squared differences relative every group's mean, so if we have m groups and n samples in each group, then `df = m*(n-1)`.

```
SSW_df = ncol(anova_mat)*(nrow(anova_mat)-1)
SSW_df
```

```
## [1] 6
```

SSB (Between Group Sum of Squares) = variation of the group mean to the overall mean. First, we find the sum of squared differences for each group mean compared to the overall mean. We also multiply by the number of values in the group to create a SS comparison for each of the original datapoints.

$$SSB = \sum_{j=1}^{m} n_j(\bar{X}_j - \bar{X})^2$$

```
SSB = 0
for ( i in 1:length(anova_mat_col_mean)) {
 SSB = SSB + (nrow(anova_mat)*(anova_mat_col_mean[i]-anova_mat_mean)^2)
}
SSB
```

```
##  a
## 24
```

For calculating between group degree of freedom, remember that if we have ***m*** groups, so it is simply ***m-1***.

```
SSB_df = ncol(anova_mat)-1
SSB_df
```

```
## [1] 2
```

F-statistic and p-value

Finally since our variance calculations are sums of squares, they can be considered to follow a χ^2 distribution. If the variance within the groups is the same and if the means of the groups are the same, then the variance between the groups should be the same as within the groups.

We can take this one step further and say that if the variance *between* the groups is *greater* than *within* the groups, then the means of the groups are different. Any change in the ratio would fit an *F-distribution* and a *p*-value can be calculated.

```
# F statistic
Fstat = (SSB/SSB_df) / (SSW/SSW_df)  # (24/2) / (6/6)
Fstat
```

```
##  a
## 12
```

```
# p-value - note that df(between) comes before df(within)
pf(Fstat, 2, 6, lower.tail = F)  # df1 = df(B) = 2; df2 = df(W) = 6
```

```
##     a
## 0.008
```

aov function

We can confirm our results using the **aov** function.

```
library(reshape2)

# we use the melt function to reshape the data frame into three columns:
# Var1 = the three groups, indexed as 1, 2, 3
# Var2 = the three groups, indexed by their variable name
# value = the value of each data point
anova_mat.melt = melt(anova_mat)
anova_mat.melt  # look at this new data structure
```

```
##   Var1 Var2 value
## 1    1    a     3
## 2    2    a     2
## 3    3    a     1
## 4    1    b     5
## 5    2    b     3
## 6    3    b     4
## 7    1    c     5
## 8    2    c     6
## 9    3    c     7
```

```
# look at the result of the ANOVA command `aov`
# the syntax is to do the analysis of the values in response to the factors (groups a,b,c)
aov(anova_mat.melt$value ~ anova_mat.melt$Var2)
```

```
## Call:
##    aov(formula = anova_mat.melt$value ~ anova_mat.melt$Var2)
##
## Terms:
##                 anova_mat.melt$Var2 Residuals
## Sum of Squares                   24         6
## Deg. of Freedom                   2         6
##
## Residual standard error: 1
## Estimated effects may be unbalanced
```

```
# summary of the aov model with F-stat and p-value
summary(aov(anova_mat.melt$value ~ anova_mat.melt$Var2))
```

```
##                     Df Sum Sq Mean Sq F value Pr(>F)
## anova_mat.melt$Var2  2     24      12      12  0.008 **
## Residuals            6      6       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA with linear model

Alternatively, we can make a linear model using `lm()` and then view the results using `anova()`.

```
anova_lm = lm(value ~ Var2, data = anova_mat.melt)
anova(anova_lm)
```

```
## Analysis of Variance Table
##
## Response: value
##           Df Sum Sq Mean Sq F value Pr(>F)
## Var2       2     24      12      12  0.008 **
## Residuals  6      6       1
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Here we can see that the overall model is significant, and the values for the sums of squares, mean sums of squares, F-statistic, and p-value are exactly the same as the values we computed by hand.

With the linear model, when applying `summary()` to the linear model allows us to look at the individual contributions from the groups.

```
summary(anova_lm)
```

```
##
## Call:
## lm(formula = value ~ Var2, data = anova_mat.melt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -1     -1      0      1      1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0000     0.5774   3.464  0.01340 *
## Var2b         2.0000     0.8165   2.449  0.04983 *
## Var2c         4.0000     0.8165   4.899  0.00271 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1 on 6 degrees of freedom
## Multiple R-squared:    0.8,  Adjusted R-squared:  0.7333
## F-statistic:    12 on 2 and 6 DF,  p-value: 0.008
```

Here we can see that groups b and c are evaluated for their contributions to the model with respect to group a, which is treated as the reference, or control sample. Both groups appear to contribute to the overall model, though group c seems more significant.

Tukey’s HSD

If we have an unplanned experiment, in which we do not have a control, we can apply Tukey’s Honest Significant Differences (Tukey’s HSD) Test to an `aov()` model to look at all pairwise differences between the groups:

```
TukeyHSD(aov(anova_mat.melt$value ~ anova_mat.melt$Var2))
```

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

```
## Fit: aov(formula = anova_mat.melt$value ~ anova_mat.melt$Var2)
##
## $‘anova_mat.melt$Var2‘
##     diff        lwr      upr     p adj
## b-a    2 -0.5052356 4.505236 0.1088670
## c-a    4  1.4947644 6.505236 0.0064937
## c-b    2 -0.5052356 4.505236 0.1088670
```

Now we can see that groups a and c are significantly different from each other, but group b (which is in the middle) is not significantly different from either of the other groups.

Unequal variances

When we did t-tests, we saw that we could use Welch's t-test when the variances between groups are not the same. There is also a Welch's version of ANOVA that we can use when the variances between groups differ.

```
oneway.test(value ~ Var2, data = anova_mat.melt, var.equal = FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  value and Var2
## F = 10.286, num df = 2, denom df = 4, p-value = 0.0265
```

Non-normal data with unequal variances

If the data are neither normally distributed nor do they have approximately equal variances, two other options are available: a Kruskal-Wallace test and a pairwise Wilcoxon Rank Sum test using p-value adjustment for multiple hypothesis testing. Options are Bonferroni (which controls for the family-wise error, FWR) correction, which is the most conservative, or Benjamini-Hochberg (which controls for the false discovery rate, FDR).

```
kruskal.test(value ~ Var2, data = anova_mat.melt)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  value by Var2
## Kruskal-Wallis chi-squared = 6.5311, df = 2, p-value = 0.03818
```

```
pairwise.wilcox.test(anova_mat.melt$value, anova_mat.melt$Var2,
                     p.adjust.method = "BH")
```

```
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute
## exact p-value with ties

## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute
## exact p-value with ties
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  anova_mat.melt$value and anova_mat.melt$Var2
##
##   a    b
## b 0.12 -
## c 0.12 0.12
##
## P value adjustment method: BH
```

Note that these will all give different results, with the tests requiring more assumptions being more restrictive, but giving greater power when the assumptions are met.