INTRODUCTION TO ANOVA

Statistical Inference

What is ANOVA?

- ANOVA is short for ANalysis Of VAriance
- Used with 3 or more groups to test for MEAN DIFFS.
- E.g., wine study with 4 groups:
  - Reference (the reference soil, the common)
  - Env 1 (the first alternative)
  - Env 2 (the second alternative)
  - Env 4 (the third alternative)
- In this case we have four levels.
- Treatment Group is people who get specific treatment or level (of the four we have in this example).
- Single factor: One Way ANOVA

Rationale for ANOVA (1)

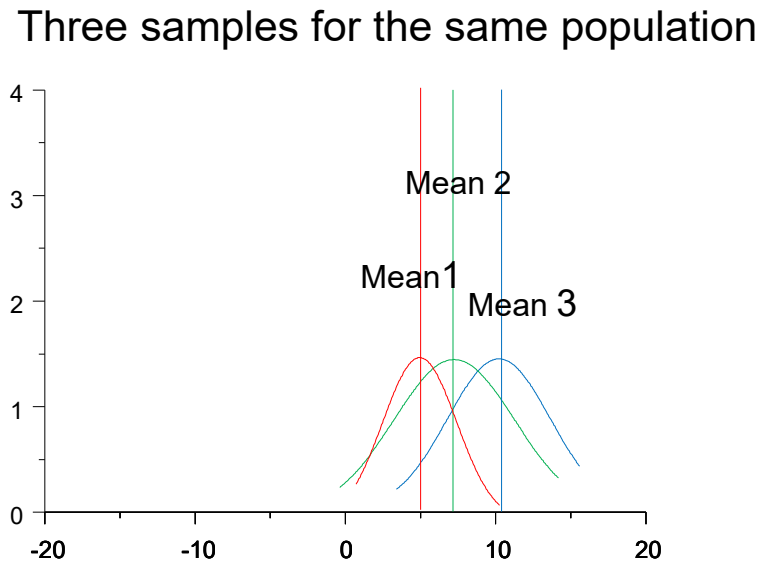
- We have at least 3 means to test, e.g., H0: $\mu 1 = \mu 2 = \mu 3$.
- Could take them 2 at a time, but really want to test all 3 (or more) at once.
- Instead of using a mean difference, we can use the variance of the group means about the grand mean over all groups.
- Logic is just the same as for the t-test. Compare the observed variance among means (observed difference in means in the t-test) to what we would expect to get by chance.

Rationale for ANOVA (2)

- HO : $\mu 1 = \mu 2 = \ldots = \mu 5$ vs. HA : not all are equal.
  - H01: $\mu 1 = \mu 2$
  - H02: $\mu 2 = \mu 3$
  - H03: $\mu 3 = \mu 4$
  - H04: $\mu 4 = \mu 5$
  - H05: $\mu 1 = \mu 3$
  - H06: $\mu 2 = \mu 4$
  - H07: $\mu 3 = \mu 5$
  - H08: $\mu 1 = \mu 4$
  - H09: $\mu 2 = \mu 5$
  - H010: $\mu 1 = \mu 5$
- Reject any null hypotheses implies reject the initial null hypothesis.
  - High computational effort.
  - Increases the type I error (reject an null hypothesis being true).

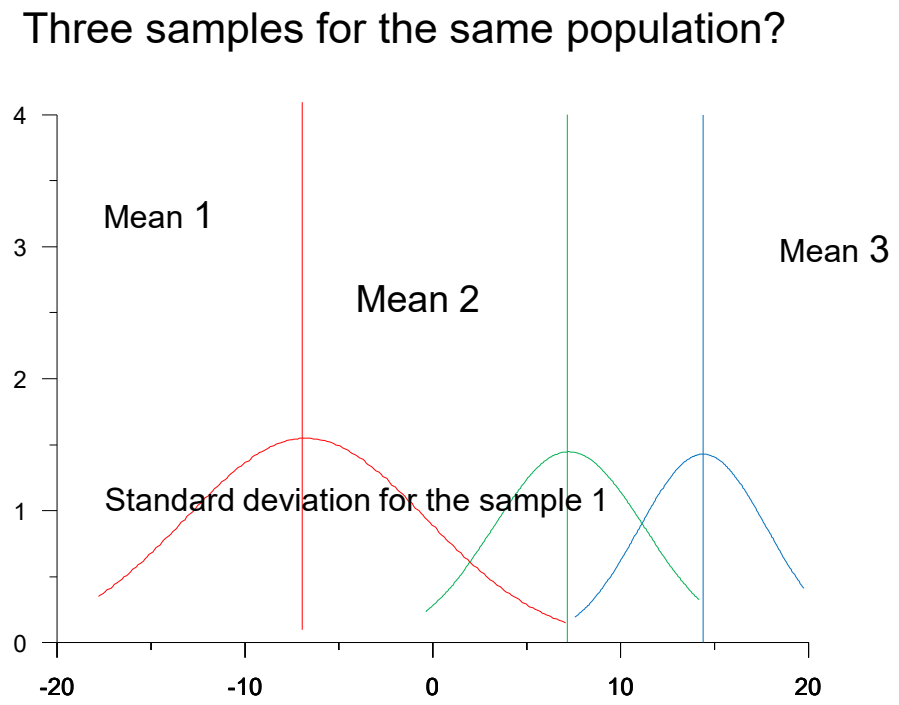
Rationale for ANOVA (3)

- Suppose we drew 3 samples from the same population.
- Note that the means from the 3 groups are not exactly the same, but they are close, so the variance among means will be small.



Rationale for ANOVA (4)

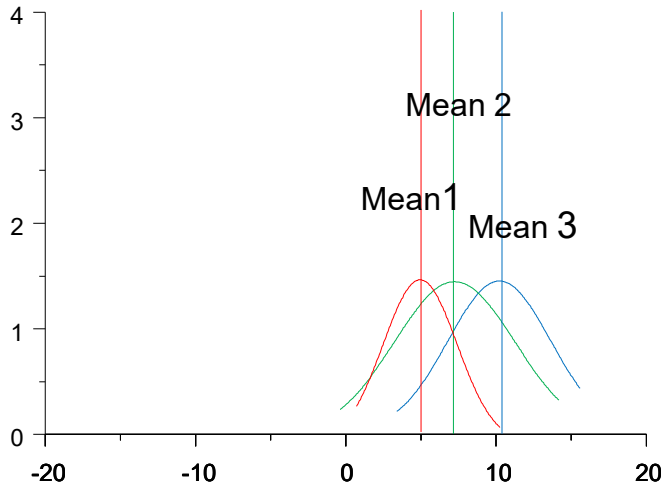
- Suppose we sample people from 3 different populations.
- Note that the sample means are far away from one another, so the variance among means will be large.



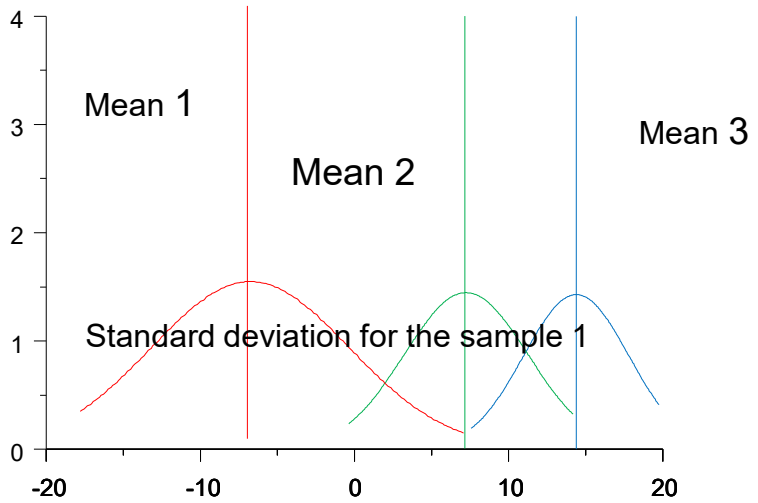
Rationale for ANOVA (5)

- Suppose we complete a study and find the following results (either graph). How would we know or decide whether there is a real effect or not?

Three samples for the same population



Three samples for different populations



- To decide, we can compare our observed variance in means to what we would expect to get on the basis of chance given no true difference in means.

Definitions

- The Grand Mean, $\bar{\bar{X}} = \bar{X}_G$, taken over all observations.
- The mean of a specific level $\bar{X}_{A_1}$ (level 1 in this case).
- The observation of the ith element X_i.

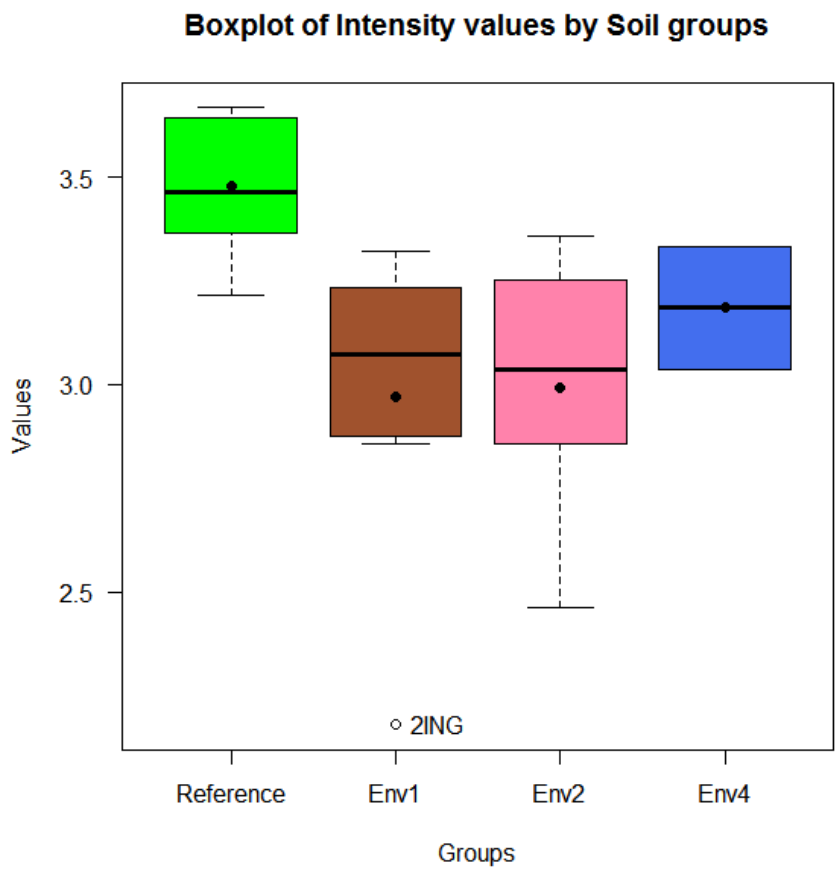
Example:
Is important the floor for the wine?

- The dataset

wine

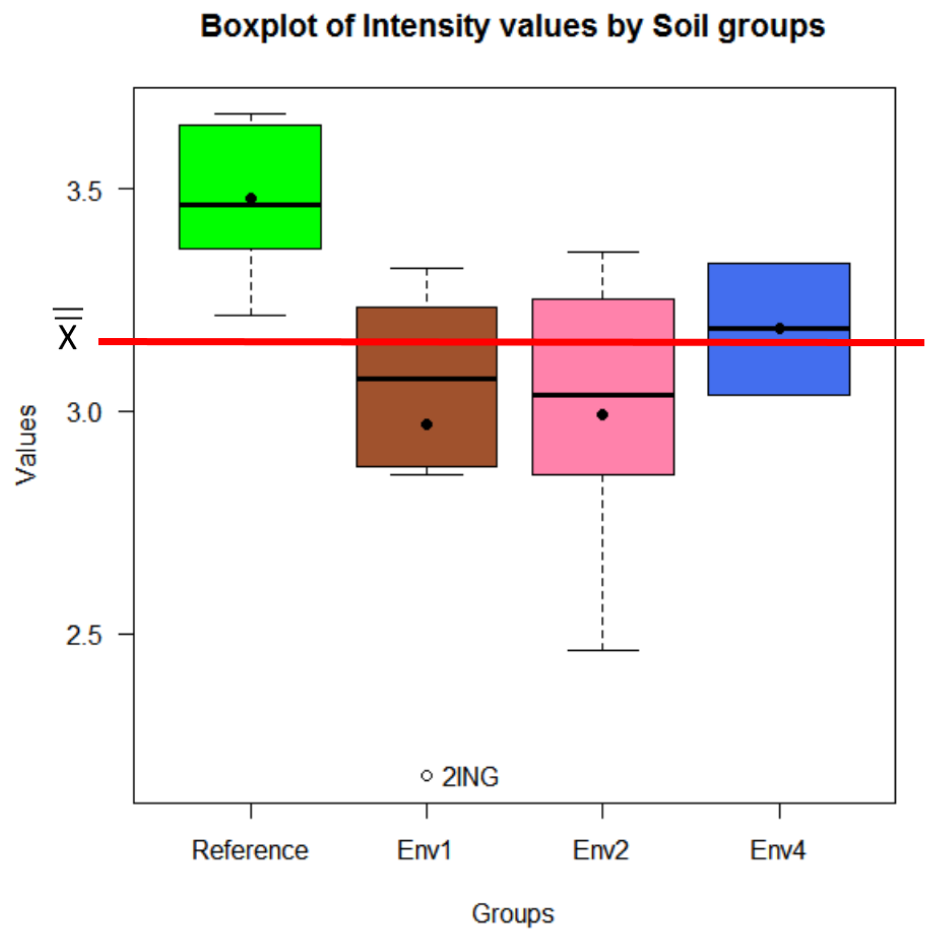
	Label	Soil	Odor.Intensity.before.shaking	Aroma.quality.before.shaking	Fruity.before.shaking	Flower.before.shaking	Spice.before.shak
2EL	Saumur	Env1	3.074	3.000	2.714	2.280	1.
1CHA	Saumur	Env1	2.964	2.821	2.375	2.280	1.
1FON	Bourgueuil	Env1	2.857	2.929	2.560	1.960	2.
1VAU	Chinon	Env2	2.808	2.593	2.417	1.913	2.
1DAM	Saumur	Reference	3.607	3.429	3.154	2.154	2.
2BOU	Bourgueuil	Reference	2.857	3.111	2.577	2.040	2.
1BOI	Bourgueuil	Reference	3.214	3.222	2.962	2.115	2.
3EL	Saumur	Env1	3.120	2.852	2.500	2.200	2.
DOM1	Chinon	Env1	2.857	2.815	2.808	1.923	2.
1TUR	Saumur	Env2	2.893	3.000	2.571	1.846	1.
4EL	Saumur	Env2	3.250	3.286	2.714	1.926	1.
PER1	Saumur	Env2	3.393	3.179	2.769	2.038	1.
2DAM	Saumur	Reference	3.179	3.286	2.778	2.231	1.
1POY	Saumur	Reference	3.071	3.107	2.731	2.120	1.
1ING	Bourgueuil	Env1	3.107	3.143	2.846	2.185	1.
1BEN	Bourgueuil	Reference	2.929	3.179	2.852	2.000	2.
2BEA	Chinon	Reference	3.036	3.179	3.037	2.231	1.
1ROC	Chinon	Env2	3.071	2.926	2.741	2.000	1.
2ING	Bourgueuil	Env1	2.643	2.786	2.536	1.889	1.
T1	Saumur	Env4	3.696	3.192	2.833	1.826	2.
T2	Saumur	Env4	3.708	2.926	2.520	2.040	2.

Example:
Is important the floor for the wine?



- We obtain information from different wines.
- There are different features (factors) that can determine the wine quality.
- We want to analyze the intensity feature.
- The factor is the floor, with 4 different possible values (levels).

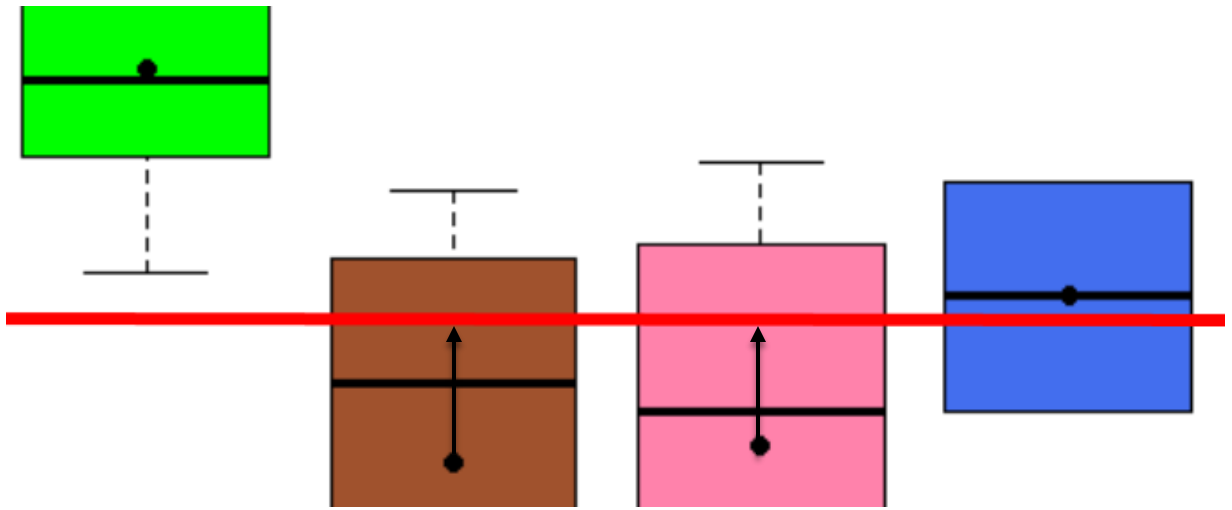
Example:
Is important the floor for the wine?



- The overall mean of the entire sample was 3.166
- This is called the "grand" mean, and is often denoted by $\bar{\bar{X}}$.
- If H0 were true then we'd expect the group means to be close to the grand mean.

Example:
Is important the floor for the wine?

- The ANOVA test is based on the combined distances from $\bar{\bar{X}}$.
- If the combined distances are large, that indicates we should reject H0.



The ANOVA statistic

SSB, MSE

The Anova Statistic

- To combine the differences from the grand mean we
  - Square the differences
  - Multiply by the numbers of observations in the groups
  - Sum over the groups
- "SSB" = Sum of Squares Between groups

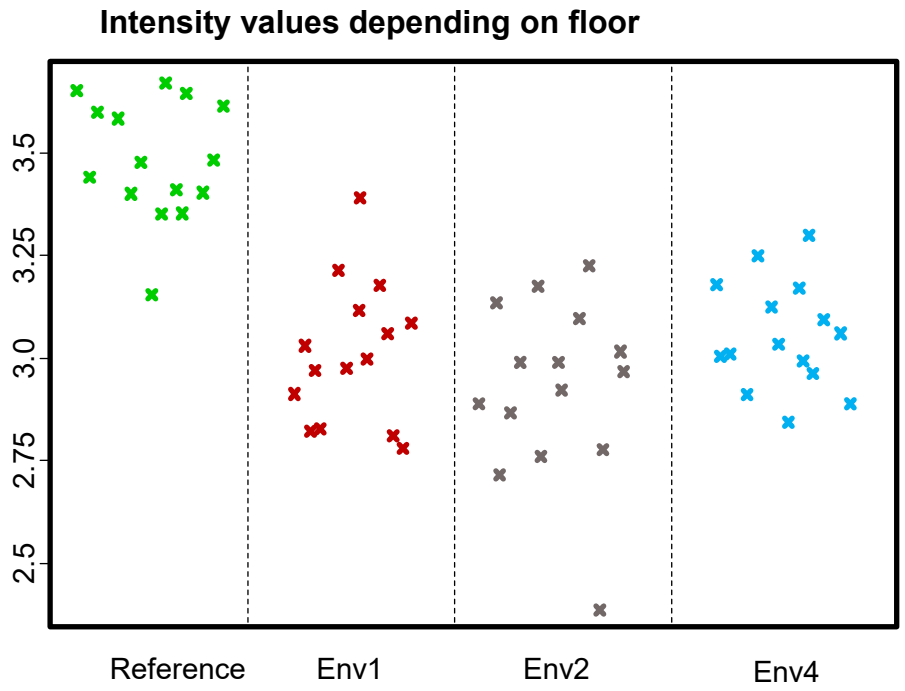
$$\mathrm{SSB} = N_{\mathrm{Re}\,ference}\left(\overline{X}_{\mathrm{Re}\,fenrenc} - \overline{\overline{X}}\right)^2 + N_{Env1}\left(\overline{X}_{Env1} - \overline{\overline{X}}\right)^2 + N_{Env2}\left(\overline{X}_{Env2} - \overline{\overline{X}}\right)^2 + N_{Env4}\left(\overline{X}_{Env4} - \overline{\overline{X}}\right)^2$$

- Where the $\overline{X}_*$ are the group means.
- Note: This looks a bit like a variance.

How big is to reject?

- For the wine data, SSB = 1.108
- Is that big enough to reject H0?
- As with the t test, we compare the statistic to the variability of the individual observations.
- In ANOVA the variability is estimated by the Mean Square Error, or MSE

MSE: Mean Square Error

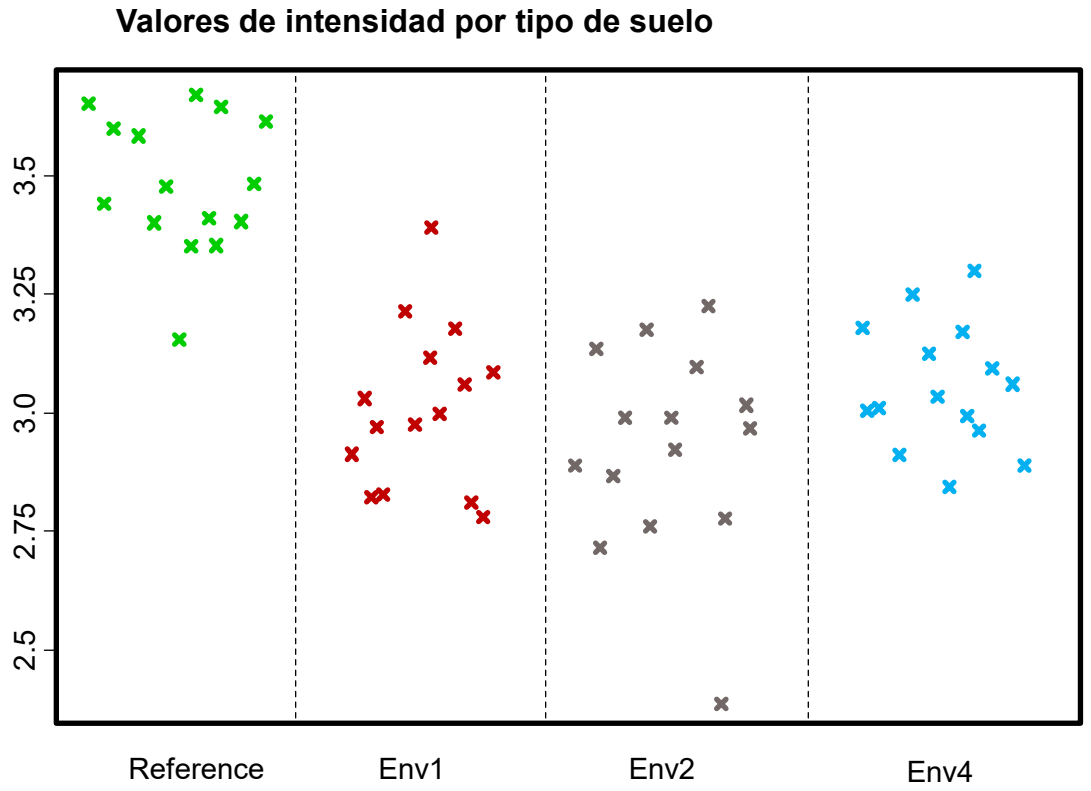


- The Mean Square Error is a measure of the variability after the group effects have been taken into account.

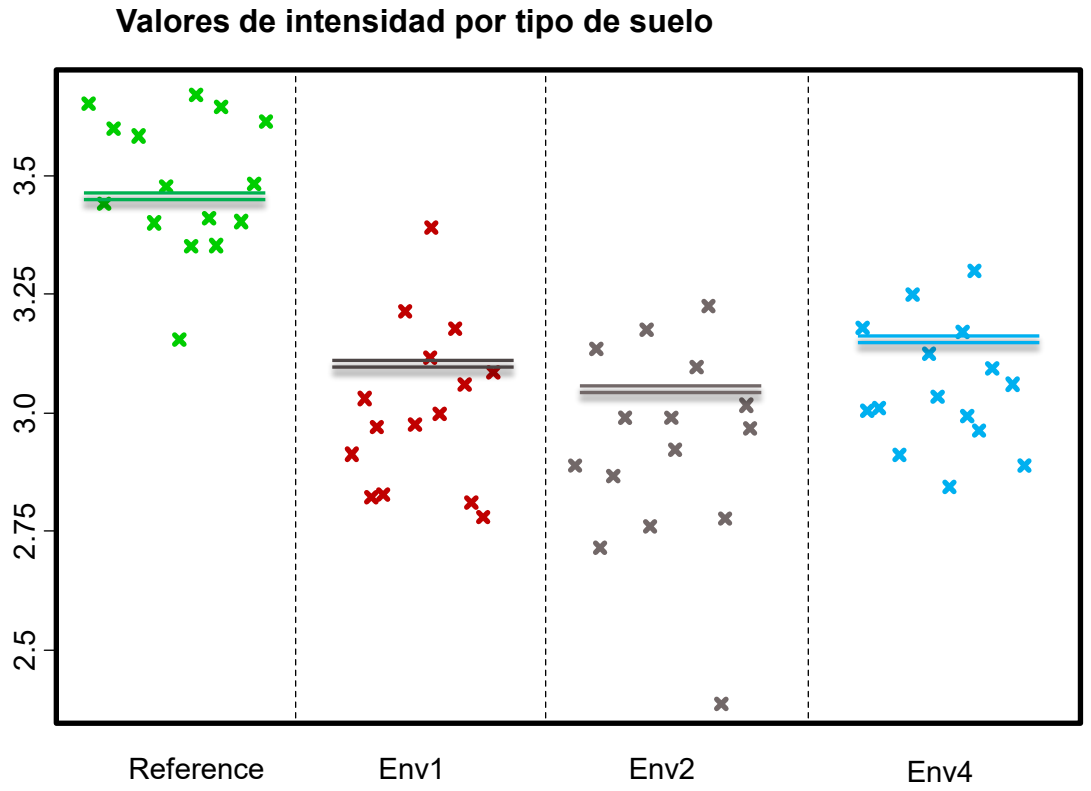
$$MSE = \frac{1}{N-K}\sum_j\sum_i\left(x_{ij} - \bar{X}_j\right)^2$$

- where xij is the ith observation in the jth group.

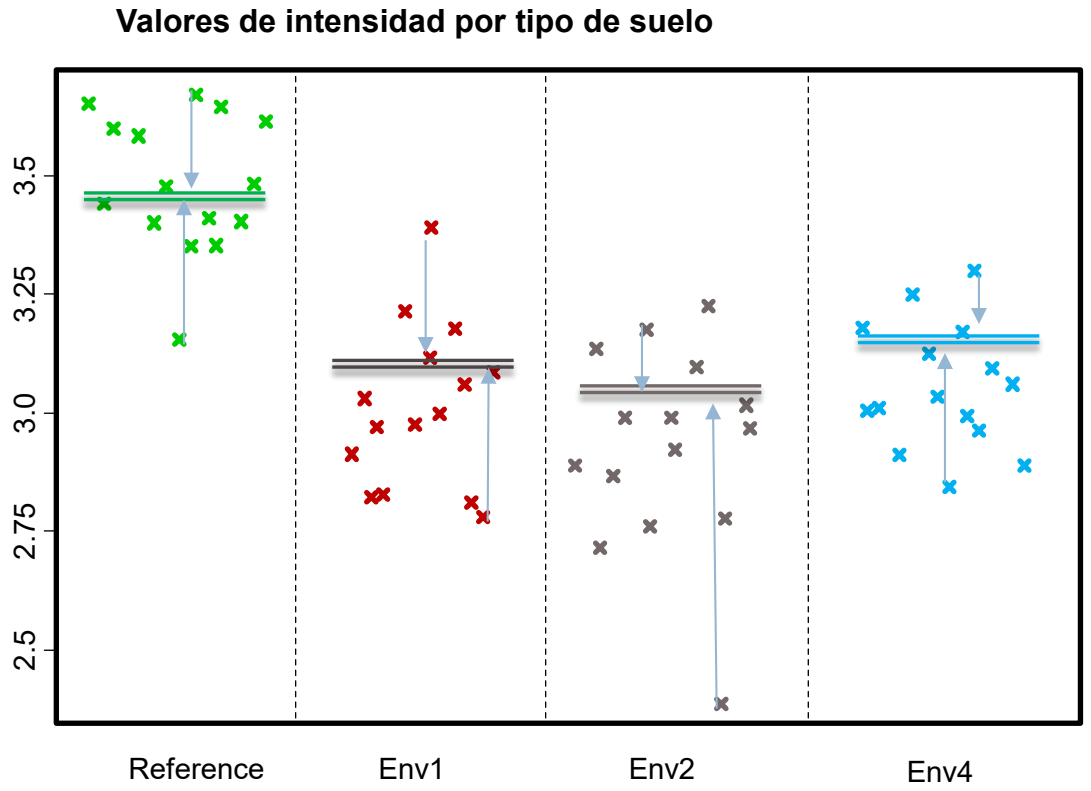
MSE: Mean Square Error



MSE: Mean Square Error



MSE: Mean Square Error



We can break the total variance in a study into meaningful pieces that correspond to treatment effects and error. That's why we call this Analysis of Variance.

Notes on MSE

- If there are only two groups, the MSE is equal to the pooled estimate of variance used in the equal-variance t test.
- ANOVA assumes that all the group variances are equal.
- Other options should be considered if group variances differ by a factor of 2 or more.

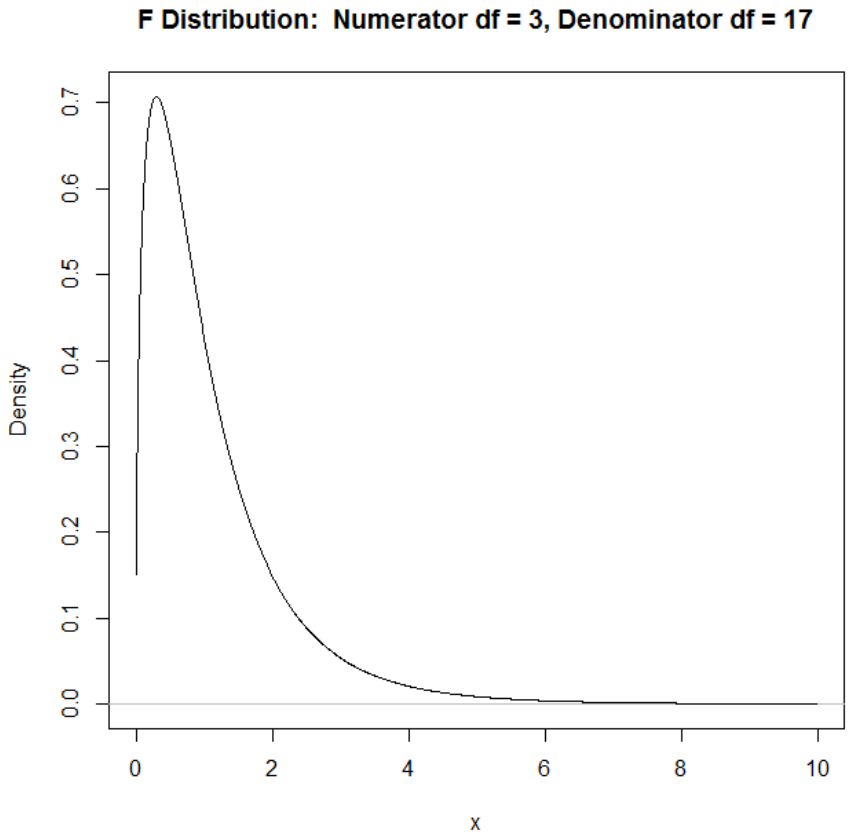
ANOVA F Test

- The ANOVA F test is based on the F statistic

$$F = \frac{SSB/(K-1)}{MSE}$$

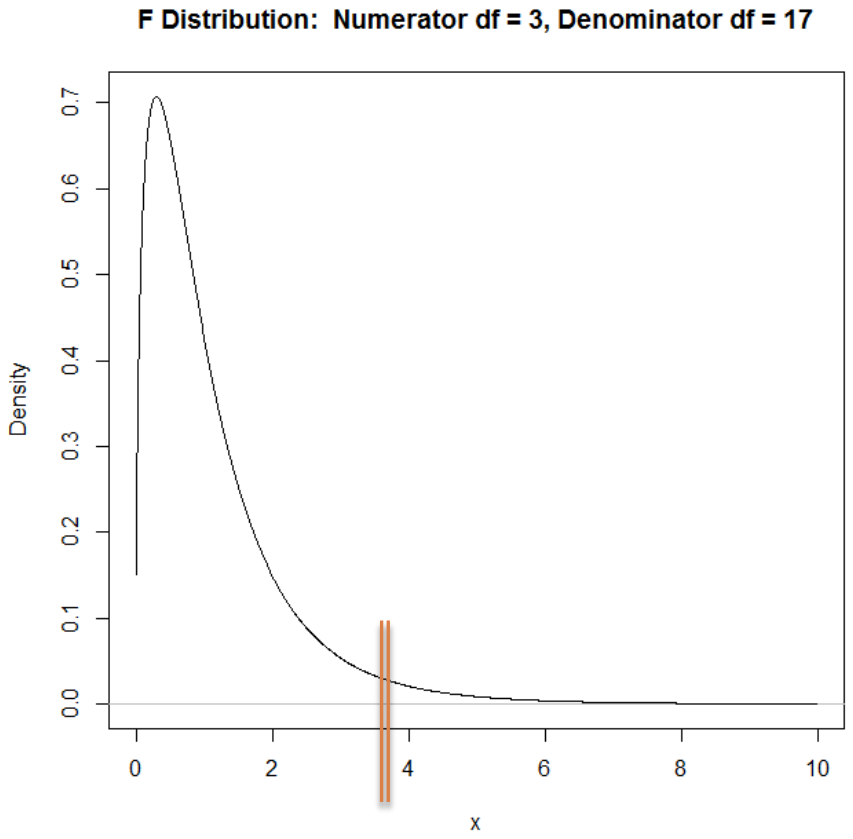
  - where K is the number of groups.
  - N is the total number of observations
- Under H0 the F statistic has an "F" distribution, with K-1 and N-K degrees of freedom (N is the total number of observations)

Wine Data: F test p-value



- To get a p-value we compare our F statistic to an F(3, 17) distribution.

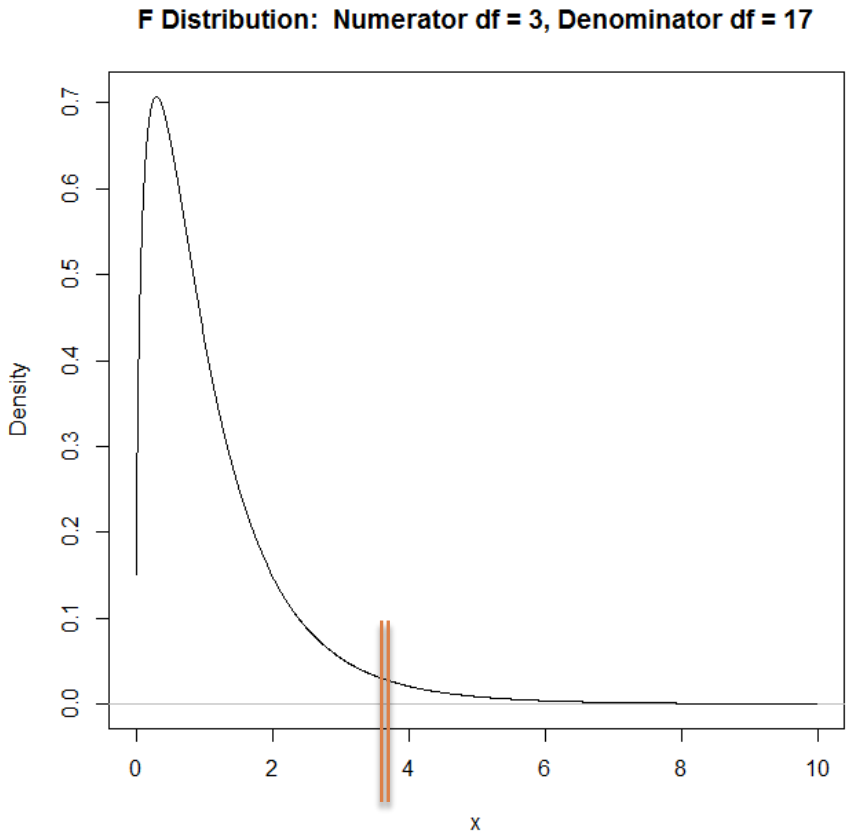
Wine Data: F test p-value



- To get a p-value we compare our F statistic to an F(3, 17) distribution.
- In our example

$$F = \frac{1.108/3}{0.0981} = 3.766$$

Wine Data: F test p-value



- To get a p-value we compare our F statistic to an F(3, 17) distribution.
- In our example

$$F = \frac{1.108/3}{0.0981} = 3.766$$

$$\mathrm{P}(\mathrm{F}(3,17) > 3.766) = 0.0306$$

- The p-value is 0.0306
- We reject H0

One-way ANOVA table

Source of variation	df	SS	MSS=SS/df	F-ration
Treatments	r-1	SSTr	MSSTr=SSTr/df	MSSTr/MSSE
Within error	N-r	SSE	MSSE=SSE/df	
Total	N-1	SST		

r: # of rows (treatments); df: degree of freedom
SS: Sum of Squares; MSS: Mean Sum of Squares

With the correction factor (One Way)

- Step 1, compute the correction factor $C = \frac{G^2}{N}$
- Step 2, SS Total: $\sum\sum\sum x_{ijk}^2 - C$ o $\sum\sum x_{ij}^2 - C$
- Step 1: SS Between Group: $\sum \frac{T_i^2}{n_i} - C$
- Step 3: SS Within Groups (error): SS Total – SS Between Group

$$SST = \sum\sum x_{ij}^2 - \frac{G^2}{N}$$

$$SS_A = \frac{(a)^2}{2^2 n}$$

$$SS_B = \frac{(b)^2}{2^2 n}$$

$$SS_{AB} = \frac{(ab)^2}{2^2 n}$$

Group	Observations				Total
1	X11	X12			T1
2	X21	X22			T2
k	xk1	xk2			
	GRAND TOTAL				$G = \sum T_k$

Tabla de ANOVA

- Results are often displayed using an ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Intra groups	3	1.108	0.3694	3.766	0.0306 *
Between groups	17	1.668	0.0981		

Sum of Squares Between (*SSB*)

Mean Square Error (*MSE*)

F Statistic

p value

R ANOVA table

- Rcmdr> AnovaModel.2 <- aov(Intensity ~ Soil, data=wine)
- Rcmdr> summary(AnovaModel.2)

```
Rcmdr>  AnovaModel.2 <- aov(Intensity ~ Soil, data=wine)

Rcmdr>  summary(AnovaModel.2)
             Df Sum Sq Mean Sq F value Pr(>F)
Soil          3  1.108  0.3694   3.766 0.0306 *
Residuals    17  1.668  0.0981
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example using R

- Continuing with the example with the Wine data, build a model that relates two independent variables such as Soil (soil where wine is grown) and label (the label of the wine). We want to analyze two dependent variables, such as intensity and aroma.

Assumptions of ANOVA

- The observations within each sample must be independent.
  - Durbin Watson test
  - dwtest(OurModel, alternative = “two.sided”)
- The populations from which the samples are selected must be normal.
  - Shapiro test
  - shapiro.test(Pop1), do for all the populations.
- The populations from which the samples are selected must have equal variances (homogeneity of variance)
  - Breusch Pagan test
  - lmtest::bptest(OurModel)

Homoscedasticity test

- Our decision rule is as follows using the 5% level of significance:
  - H0 (Null Hypothesis): Homoscedasticity
  - HA (Alternative Hypothesis): Heteroscedasticity

Homocedasticity

lmtest::bptest(AnovaModel.3)

*BP = 2.7267, df = 3, p-value = **0.4357***

We accept H0

- The recommended method for correcting heteroscedasticity is redefining the variables (ex. Log).