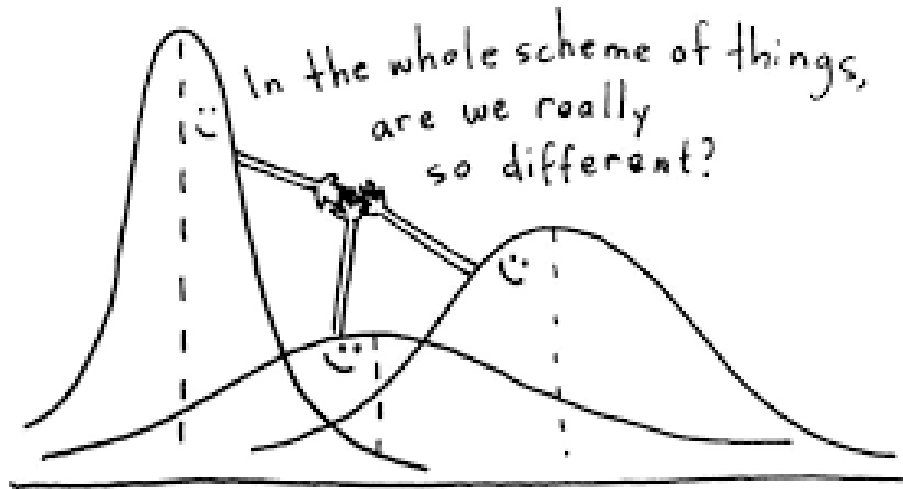# INTRODUCTION TO ANOVA

Statistical Inference

Nihan Acar-Denizli, Pau Fonseca

What is ANOVA?

- ANOVA is short for ANalysis Of Variance.

- The objective is to test the difference between means for 3 or more groups.

In the whole scheme of things, are we really so different?

What is ANOVA?

- Example: 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)
- Level is value, kind or amount of IV
- Treatment Group is people who get specific treatment or level of IV
- Treatment Effect is size of difference in means

Hypotheses of ANOVA

H_0: The population means are all equal.
$$(\mu_1 = \mu_2 = \cdots = \mu_k)$$

H_1: At least one of the population means is different.
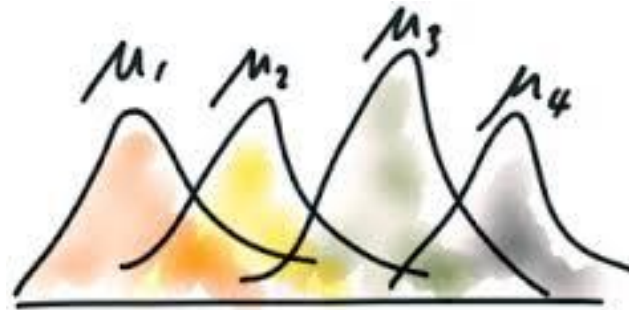
Assumptions of ANOVA

☐ For each population, the response variable is normally distributed (checked by Histogram, Q-Q plot or Shapiro-Wilk hypothesis test).

☐ The variance of the response variable is the same for all of the populations (checked by Levene's Test).

☐ The observations must be independent (checked by Durbin Watson test).

Rationale for ANOVA (1)

- We have at least 3 means to test, e.g.,

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

- 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.
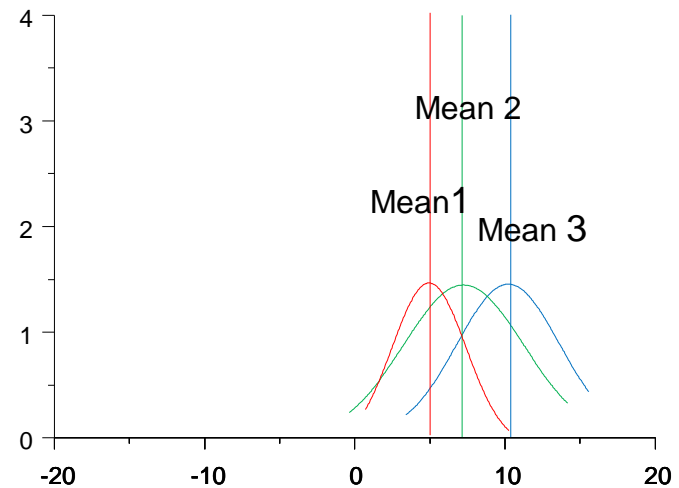
Rationale for ANOVA (2)

- 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.

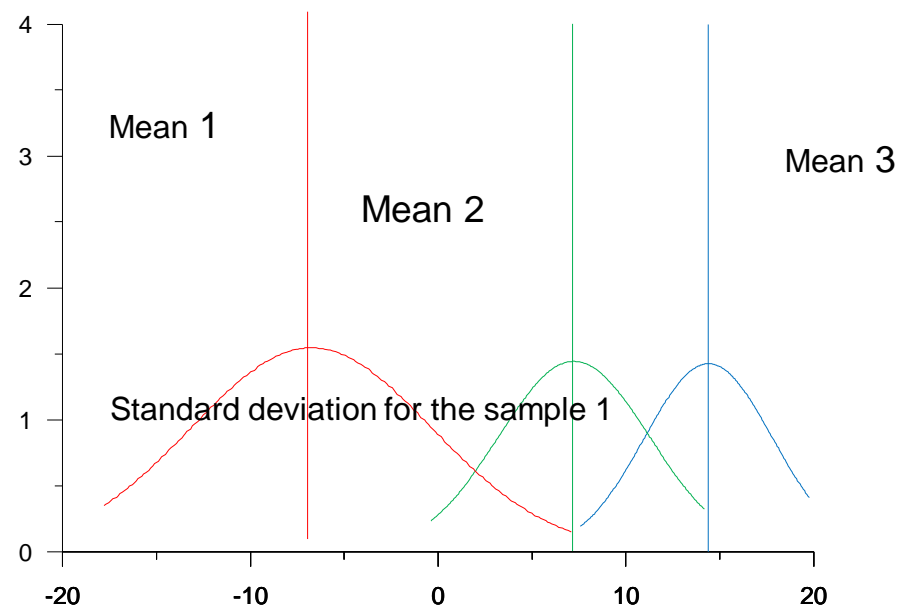Three samples for the same population

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.

Three samples for the same population?

Mean 1

Mean 2

Mean 3
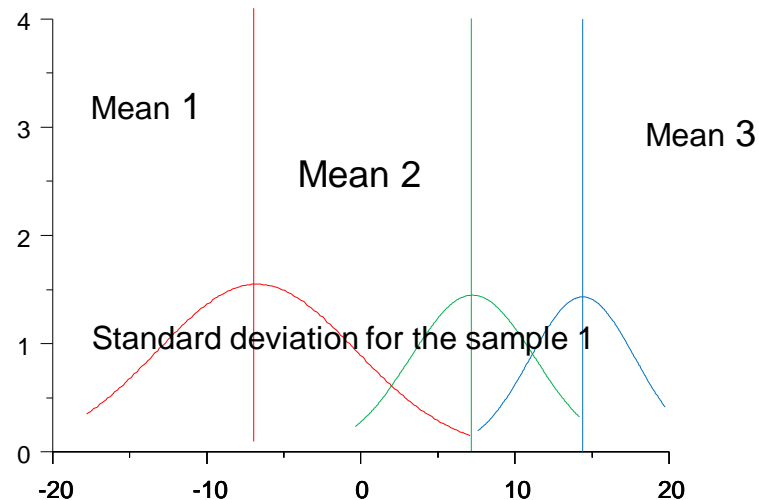
Standard deviation for the sample 1

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
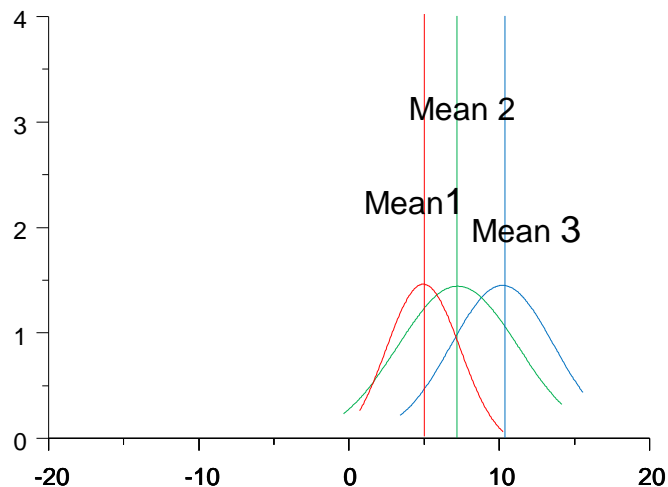
Three samples for different populations

- 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.
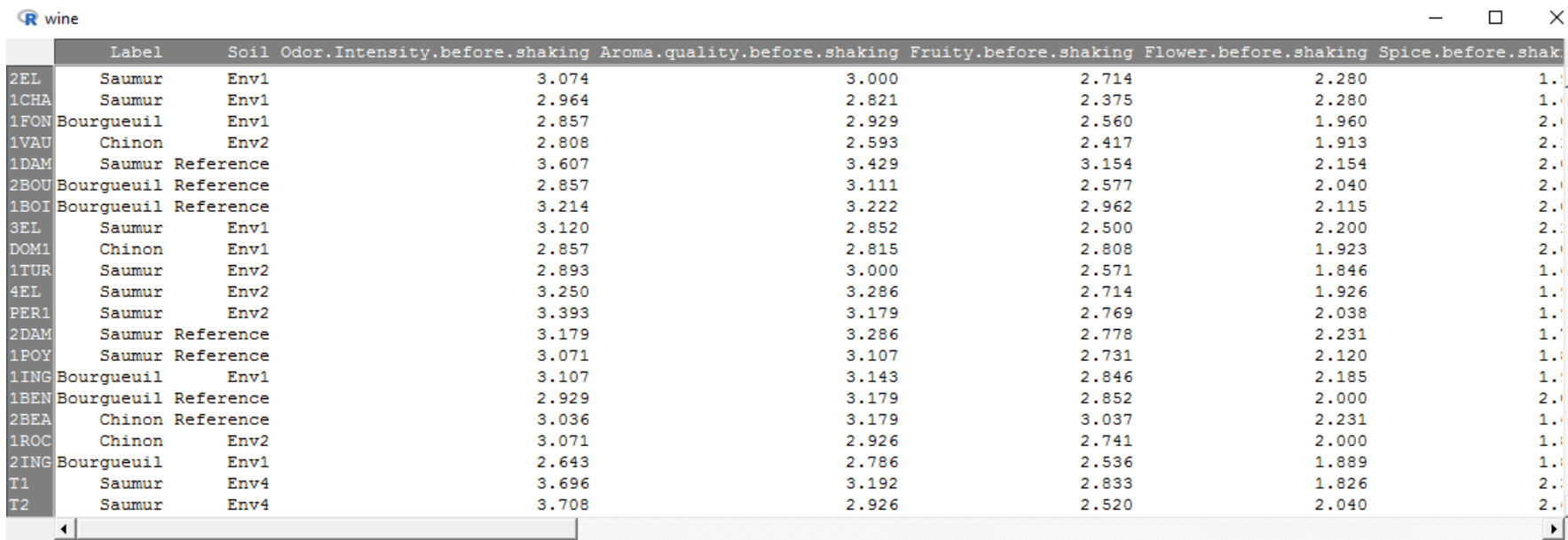
Definitions

☐ 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 Wine dataset

|	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?

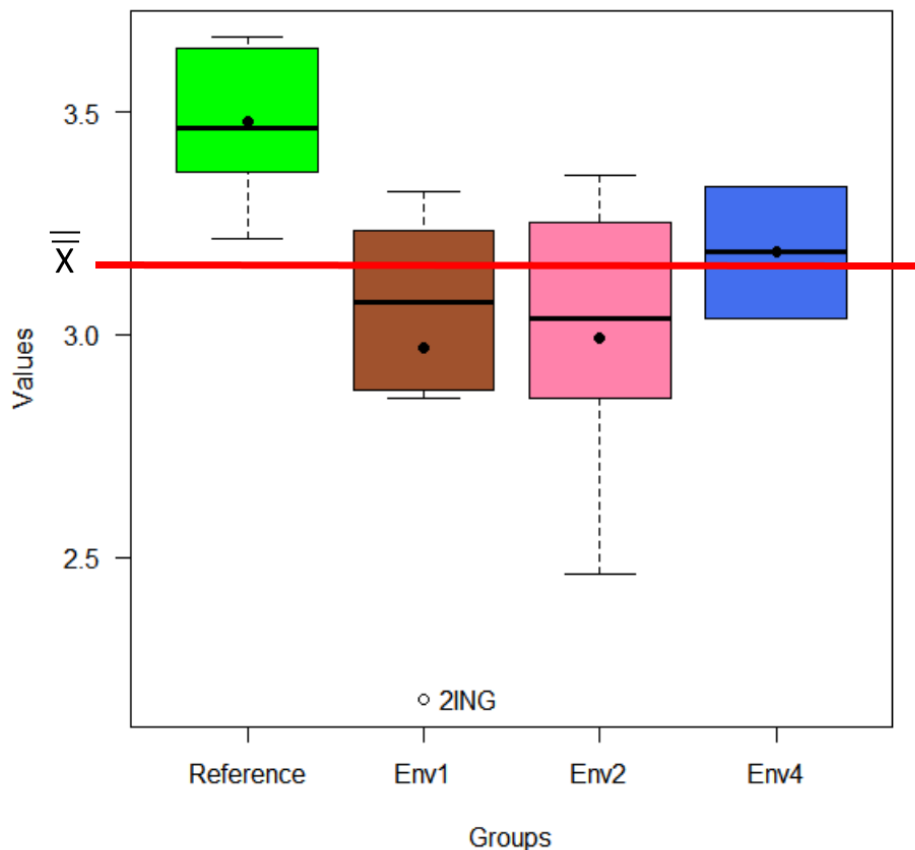

Boxplot of Intensity values by Soil groups
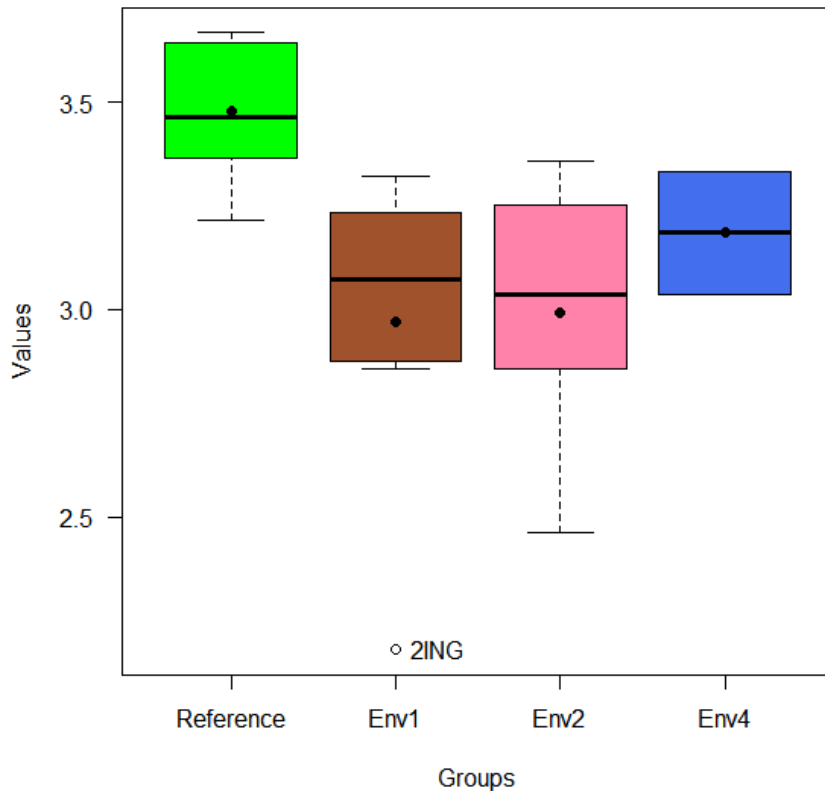
- We obtain information from different wines.

- 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?

**Boxplot of Intensity values by Soil groups**

- □ 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.
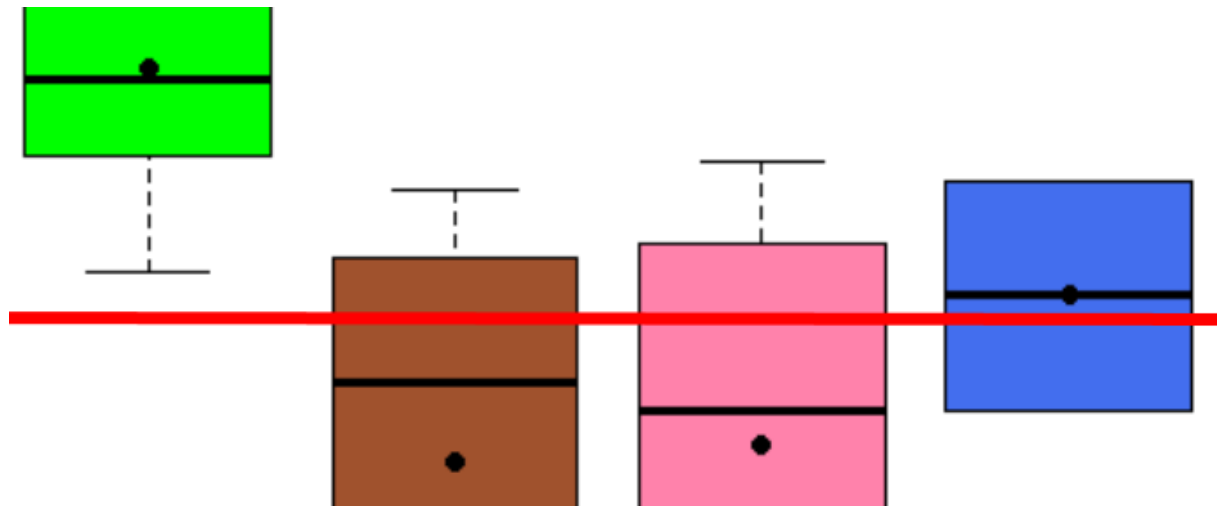
The ANOVA statistic

SSB, SSE, SST

Variation in ANOVA

- Between group variability refers the variation between the distribution of groups which is measured by looking at the differences between the overall mean and the group means (SSB).

- Within group variability refers variation caused by the differences within individual groups and is measured by looking at how each value in each sample differs from its respective sample mean (SSW or SSE).

Variation in ANOVA
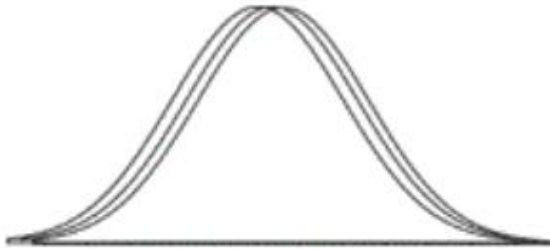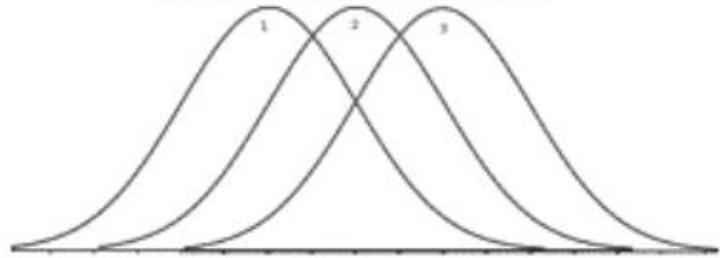
Source: Psychstat – Missouri State

Sum of Squares Between Groups (SSB)

- 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

$$\text{SSB} = \sum_{j=1}^{k} N_j \left(\overline{X}_j - \overline{\overline{X}} \right)^2$$

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

where the $\overline{X}_*$ are the group means.

Sum of Squares Within Groups (SSE)

- "SSE" = Sum of Square Error (Sum of Squares Within Groups)

- Sum of Squares computed as the difference between the observations and group means. This indicates error.

$$SSE = \sum_{i=1}^{n_j} \sum_{j=1}^{k} \left(X_{ij} - \overline{X}_j \right)^2$$

Total Sum of Squares (SST)

- The sum of squares computed by the difference between observed values and the grand mean.

$$SST = \sum_{i=1}^{n_j} \sum_{j=1}^{c} \left(X_{ij} - \bar{\bar{X}} \right)^2$$

$$SST = SSB + SSE$$

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 computed by the division of SSE to its degree of freedom.

□ The degrees of freedom of SSE is equal to the difference between the number of observations and the number of groups, which is denoted by N-k.

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

MSE: Mean Square Error

Intensity values depending on floor
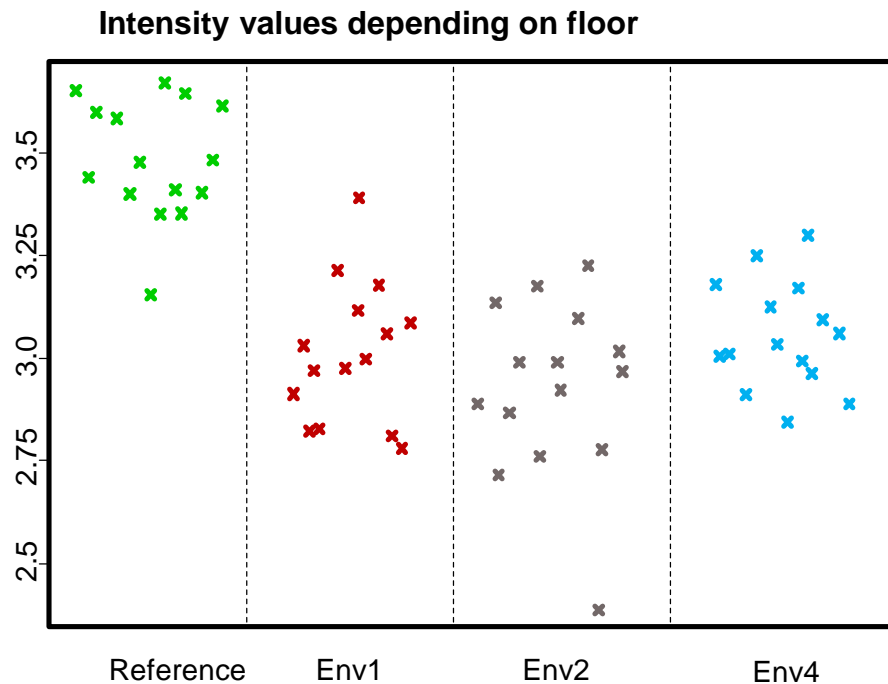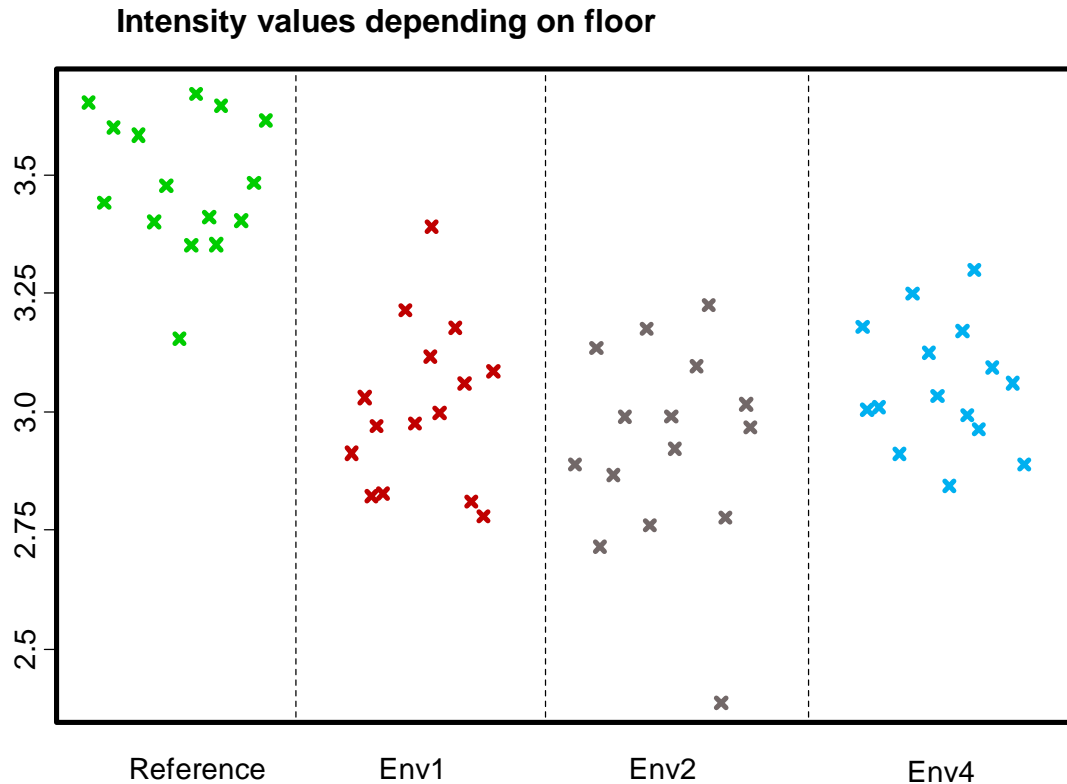

☐ The Mean Square Error is a measure of the variability after the group effects have been taken into account.

$$MSE = \frac{1}{N-k}\sum_{j}\sum_{i}\left(X_{ij} - \overline{X}_{j}\right)^{2}$$

☐ where X_{ij} is the ith observation in the jth group.

MSE: Mean Square Error

Intensity values depending on floor

MSE: Mean Square Error

Intensity values depending on floor

MSE: Mean Square Error
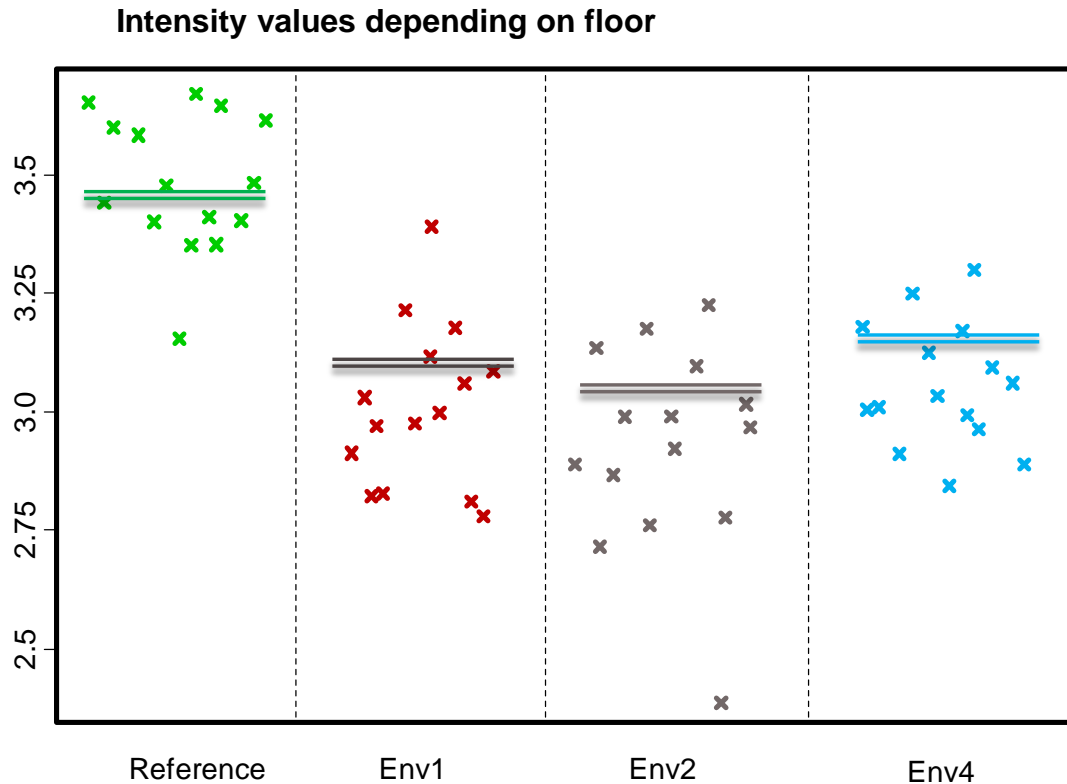
Intensity values depending on floor

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 two 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 H_0 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

F Distribution: Numerator df = 3, Denominator df = 17
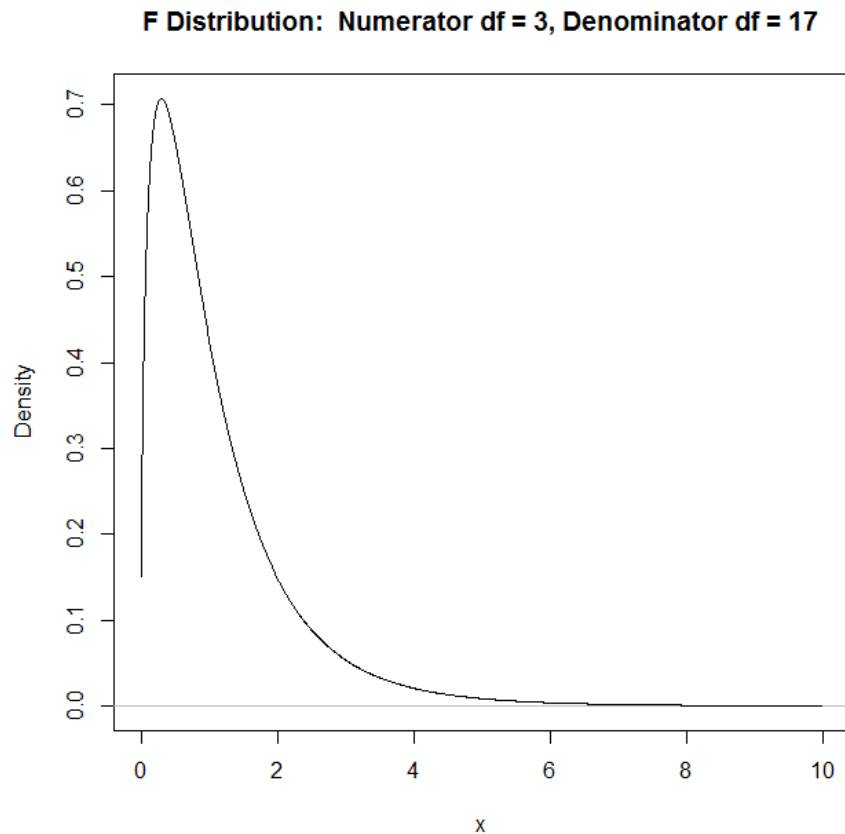
- To get a p-value we compare our F statistic to an F(3, 17) distribution.

Wine Data: F test p-value
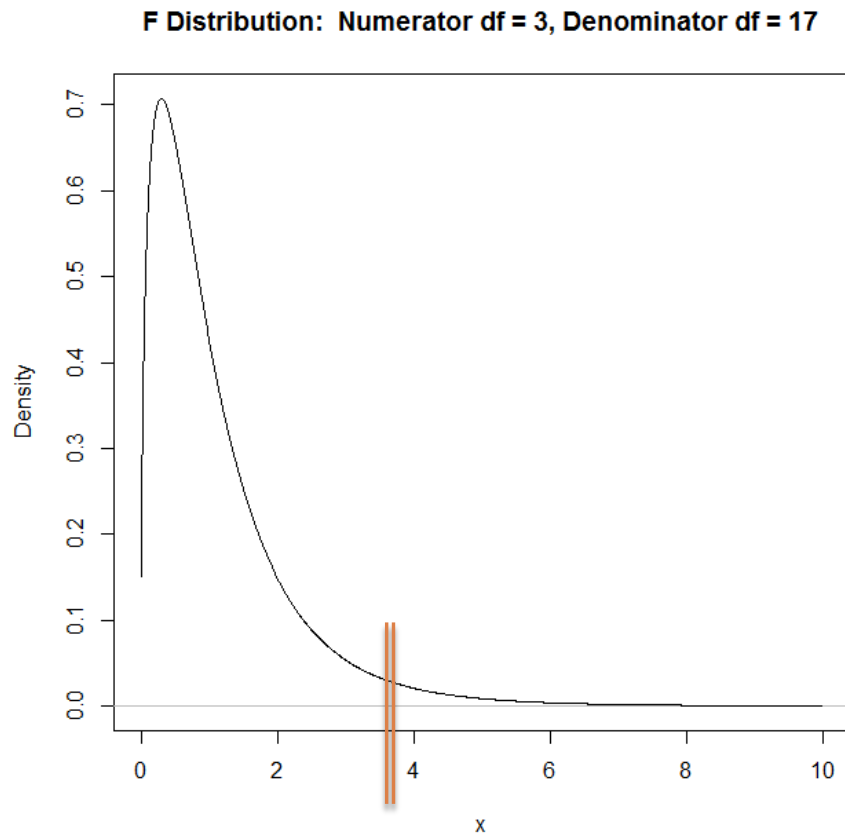
F Distribution: Numerator df = 3, Denominator df = 17

- 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
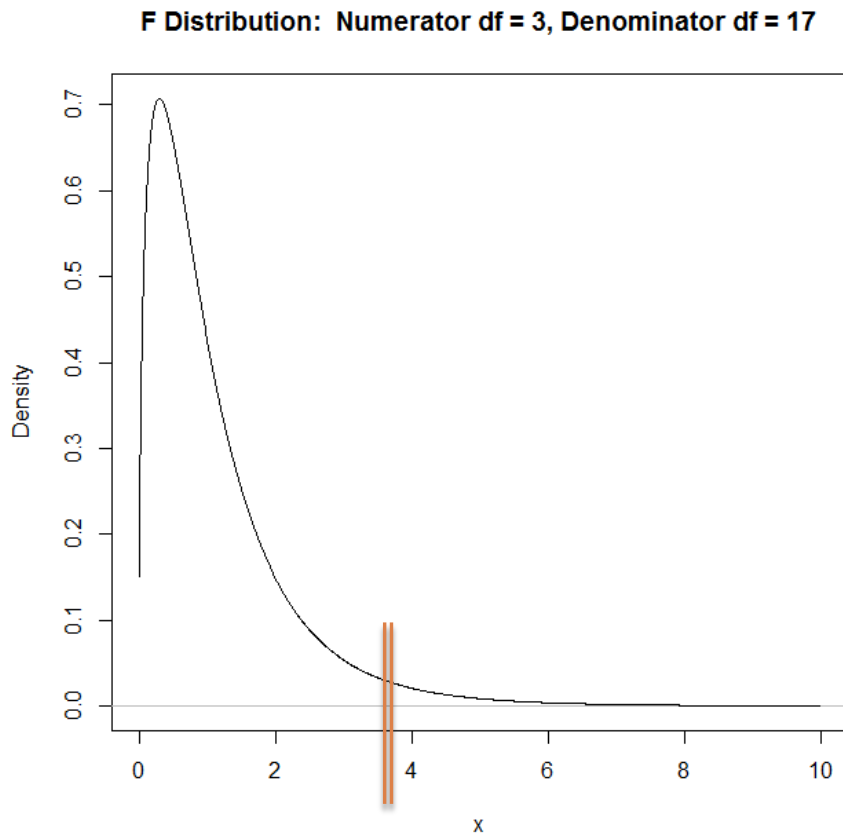
F Distribution: Numerator df = 3, Denominator df = 17

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

- In our example
$$F = \frac{1.108/3}{0.0981} = 3.766$$
$$P(F(3,17) > 3.766) = 0.0306$$

- The p-value is 0.0306
- We reject H_0

ANOVA Table

□ Results are often displayed using an ANOVA Table

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

Sum of Squares Between (*SSB*)

Mean Square Error (*MSE*)

F Statistic

p value

R ANOVA table

- AnovaModel.2 <- aov(Intensity ~ Soil, data=wine)
- 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 (Two-Way ANOVA).

Assumptions of ANOVA

- The observations within each sample must be **independent.**
  - Durbin Watson test
  - dwtest(Model, 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)
  - Levenes Test / Box's M Test / Breusch Pagan test
  - car::leveneTest(Model) / biotools::boxM(var,group) / lmtest::bptest(Model)

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).

Multiple comparisons

Post-hoc testing
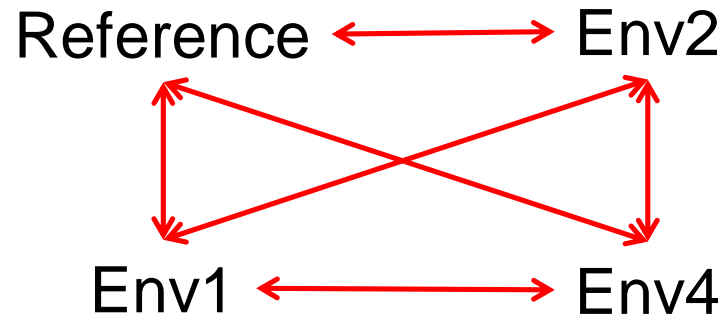
Multiple Comparisons

☐ Post-hoc testing usually involves multiple comparisons.

☐ For example, if the data contain 4 groups, then 6 different pairwise comparisons can be made

Multiple Comparisons

- Each time a hypothesis test is performed at significance level α, there is probability α of rejecting in error.

- Performing multiple tests increases the chances of rejecting in error.

The problem of the multiple comparisons

- If one test is performed at the 5% level, there is only a 5% chance of incorrectly rejecting the null hypothesis if the null hypothesis is true.

- For 100 tests where all null hypotheses are true, the expected number of incorrect rejections is 5.

The Binomial Probability Distribution

- For a binomial experiment with:
  - **n** trials and
  - probability **p** of success on a given trial, q = 1 – p.
  - the probability of **k** successes in n trials is

$$P(x = k) = C_k^n p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k} \text{ for } k = 0,1,2,...n.$$

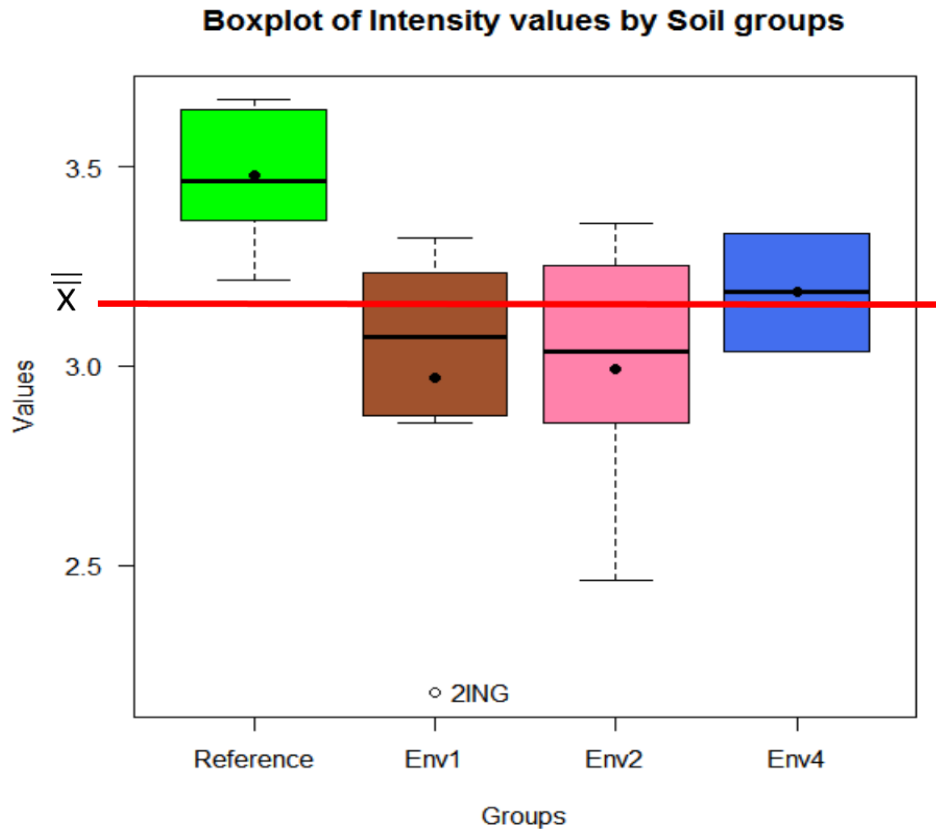Recall $\quad C_k^n = \dfrac{n!}{k!(n-k)!}$

with $n! = n(n-1)(n-2)...(2)1$ and $0! \equiv 1$.

Multiple Comparisons

☐ If the tests are independent, the probability of at least one incorrect rejection is 99.4%. These errors are called false positives or Type I errors.

☐ P(at least one rejection) = 1-P(no rejections)

☐ 1-P(x=0) = 1-0.0059 = .994

$$P(x=0) = \frac{100!}{0!(100-0)!}0.05^{0}0.95^{100-0}$$

Post-hoc Testing

Boxplot of Intensity values by Soil groups

- Analyzing the wine intensity regarding the soil.

- The ANOVA shows good evidence (p = 0.0306) that the means are not all the same.

- Which means are different?

- Can directly compare the subgroups using "post hoc" tests.

Post-hoc Testing

- If the t-test is significant, you have a difference in population means.

- If the F-test is significant, you have a difference in population means. But you don't know where.

- With 3 means, could be A=B>C or A>B>C or A>B=C.

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

Post-hoc Testing

- ANOVA just says that the means differ, but not which ones. We have to do additional tests to determine.

- When are post hoc tests done? As the name implies after an ANOVA
  - But only after a rejection of the null hypothesis.
 - Only if there are 3 or more treatments; $k > 2$. If only 2 treatments we can just do a t-test.

Post-hoc Testing

- Post hoc tests are going to let us go back through our data and compare individual treatments two at a time:
  - Bonferroni correction
  - Tukey's HSD test
  - Least Significant Difference Test
  - Duncan's new multiple range test
  - Scheffé's method
  - Dunnett's test
  - Friedman test
  - Newman–Keuls method

Example: Pairwise t test without correction

data(wine, package="FactoMineR")

pairwise.t.test(wine$Intensity, wine$Soil, p.adj="none")
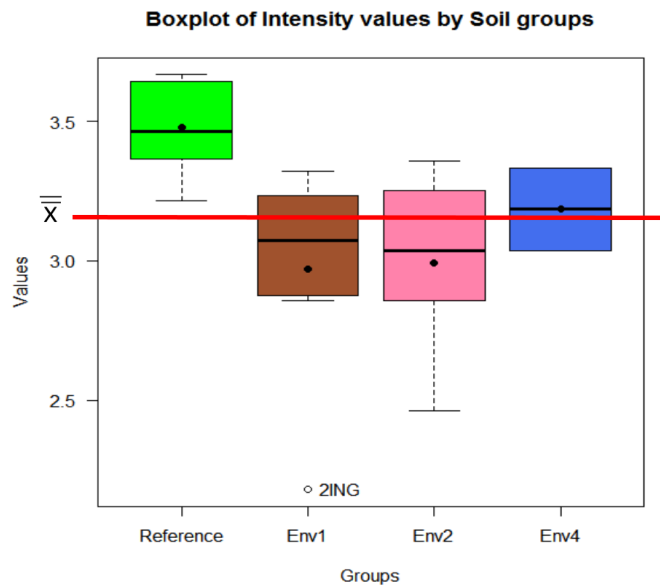
Boxplot of Intensity values by Soil groups

Pairwise comparisons using t tests with pooled SD

data: Intensity and Soil

Reference	Env1	Env2
Env1 0.0074	-	-
Env2 0.0166	0.9052	-
Env4 0.2565	0.4048	0.4728

P value adjustment method: none

Bonferroni Correction

- The Bonferroni correction is a simple way to adjust for the multiple comparisons. We calculate a new α value.
  - Perform each test at significance level α.
  - Divide the α by the number of tests performed.
    - $\alpha' = \alpha/m$

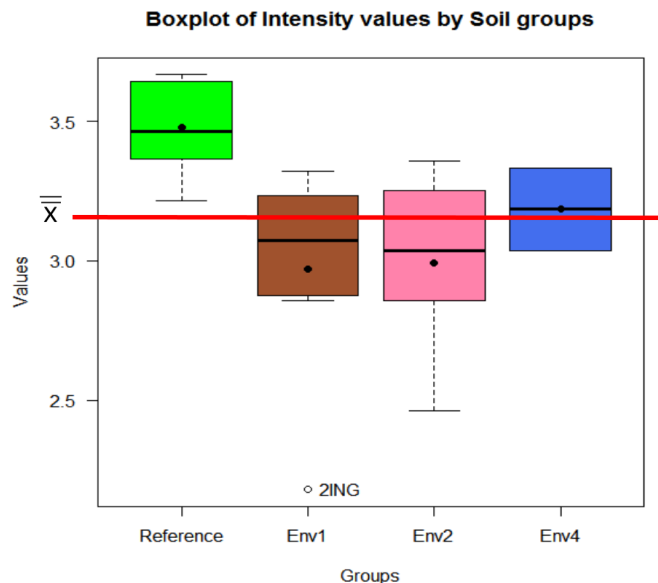- Is a conservative approach. Useful when the number of groups is small.

Example Bonferroni: wine

data(wine, package="FactoMineR")

with(wine,

{

pairwise.t.test(Intensity,Soil,p.adj="bonf")

})

Conservative with large number of tests.

Boxplot of Intensity values by Soil groups

Pairwise comparisons using t tests with pooled SD

data: Intensity and Soil

Reference	Env1	Env2
Env1 0.044	-	-
Env2 0.100	1.000	-
Env4 1.000	1.000	1.000

P value adjustment method: bonferroni

Least Significant Difference test

- The computation is very similar to the equal-variance t test.

- Compute an equal-variance t-test, but replace the pooled variance (s^2) with the *MSE*.

Least Significant Difference test

- Remember, equal variance test

$$\frac{\overline{y}_A - \overline{y}_B}{s\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}} > t_{1-\alpha,n}$$

- Remember, MSE

$$MSE = \frac{1}{N-K}\sum_{j}\sum_{i}\left(x_{ij} - \overline{X}_j\right)^2$$

$$\left|\overline{y}_{i\cdot} - \overline{y}_{j\cdot}\right| \geq t_{\alpha/2}\sqrt{s_W^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Tukey's HSD (Honest significant difference)

$$|\bar{X}_i - \bar{X}_j| \geq q(\propto, k, df)\sqrt{\frac{MSE}{n}}$$

Equal number of samples

$$|\bar{X}_i - \bar{X}_j| \geq q(\propto, k, df)\sqrt{\frac{MSE}{n}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

$$k = \#groups, df = N - k$$

Studentized range distribution

$$q = \frac{(\bar{y}_{max} - \bar{y}_{min})}{S\sqrt{2/n}}$$

Example Tukey: wine

data(wine, package="FactoMineR")

with(wine,

{

  aov_model=aov(Intensity~Soil)

  TukeyHSD(aov_model)

})

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = Intensity ~ Soil)

$Soil

> When confidence intervals are needed or sample sizes are not equal.

	diff	lwr	upr	p.adj
Env1-Reference	-0,5094286	-0,9853299	-0,03353	0,033641
Env2-Reference	-0,4872571	-1,0085809	0,034067	0,071501
Env4-Reference	-0,2948571	-1,0087091	0,418995	0,650531
Env2-Env1	0,02217143	-0,4991523	0,543495	0,999342
Env4-Env1	0,21457143	-0,4992805	0,928423	0,827774

To know more

- Part III: Stochastic Processes Chapter 11.2: Regression of **Probability and Statistics for Computer Scientists** (2014 Ed.)