STAT 212: Principles of Statistics II

Lecture Notes: Chapter 2 (Part A)

'One-Way' Analysis of Variance (ANOVA)

Sharmistha Guha

Dept. of Statistics

Texas A&M University

Fall 2022

Analysis of Variance

Analysis of variance, or *ANOVA*, is a methodology for analyzing the results of many different kinds of experiments.

Ironically, the main use of ANOVA is in comparing *means*!!

Variance enters the picture as we measure the amount of *variance between different means* across different groups as a way of deciding if the means are significantly different.

An experiment is done in which data are obtained under *two or more different sets of conditions* (determined by a 'factor' or 'factors').

The simplest ANOVA is the *single factor*, or *one-way* ANOVA of Chapter 10 (in textbook).

Single factor ANOVA:

The different sets of conditions correspond to the values (or levels) of a single factor.

Examples

1. 150 students are randomly assigned to one of three classes, 50 students per class. The same instructor is used for all three classes, but a different text is used in each one. All students take the same test at the end of the course. Do the scores differ significantly across classes? *Factor is text.*

2. A certain strain of corn is grown at a number of different locations. Four different fertilizers are used to determine which results in the largest yield of corn. *Factor is fertilizer.*

3. Three companies produce a particular type of electrical component. Random samples of components are selected from the output of each company. Components are put into service until they "die." Does average lifetime differ among the three companies? *Factor is company.*

General setup:

We have k groups or *treatments.*

Measurements are recorded for n_i subjects in group i. The measurements in group i are denoted as:

$$X_{i1}, X_{i2}, \ldots, X_{in_i},$$

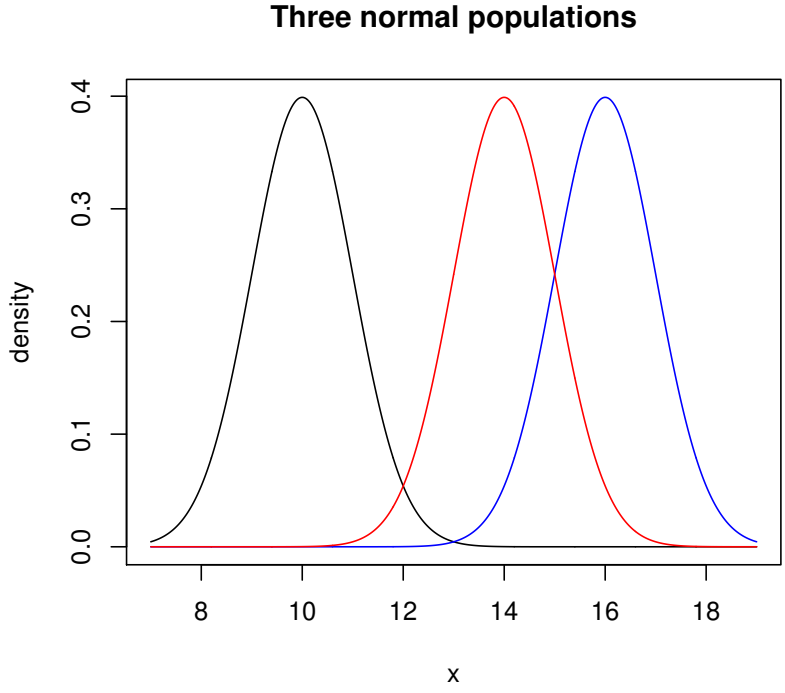
and i ranges from 1 to k.

The statistical model is:

$$X_{ij} = \mu_i + \epsilon_{ij}, \quad j = 1, \ldots, n_i; \ i = 1, \ldots, k.$$

- μ_i is the population mean for group i.
- $\epsilon_{ij} \sim N(0, \sigma^2)$ for all i and j.
- All the ϵ_{ij}s are independent of each other.

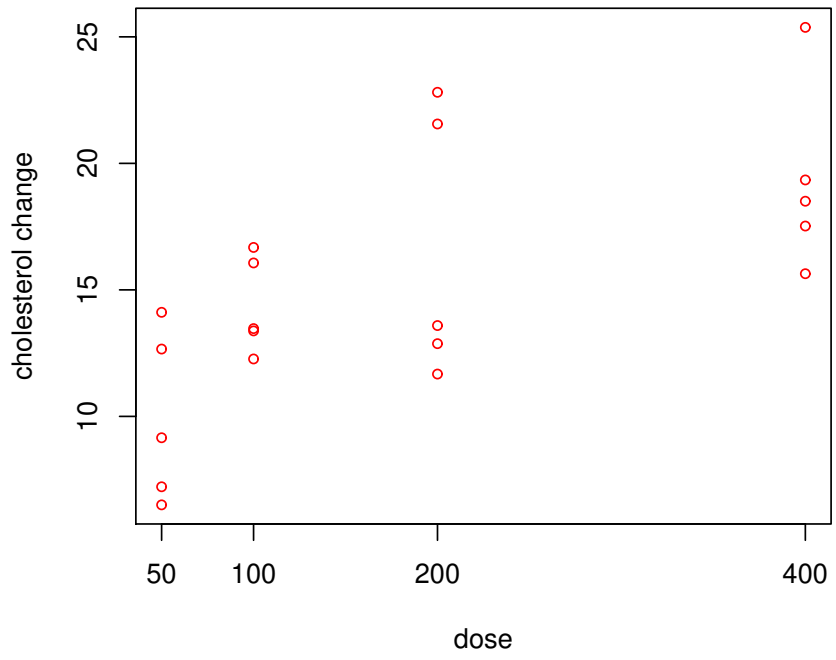
This model is equivalent to assuming that *the ith group of data, $X_{i1}, \ldots, X_{in_i}$, is a random sample from $N(\mu_i, \sigma^2)$, and that the k samples are independent of each other*.



This situation is not so different than regression! Suppose, for example, that the factor levels are values of a numerical variable.

The factor might be *dose level of a drug*. Suppose the four dose levels are 50 mg, 100 mg, 200 mg and 400 mg.

A given dose level is administered to each of 5 patients and the change in, say, cholesterol level is measured for each patient. The resulting data might look like this:



A main problem of interest in one-way ANOVA is testing the *hypothesis of equal population means:*

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k.$$

The alternative hypothesis is

$$H_a : \text{at least two } \mu_i\text{s are different.}$$

The population means μ_i's are estimated by the respective sample means:

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \quad i = 1, \ldots, k.$$

The *grand mean* is:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} = \frac{1}{N} \sum_{i=1}^{k} n_i \bar{X}_i,$$

where $N = n_1 + \cdots + n_k$.

Rationale for the test of H_0:

Compare how much variance there is *between* treatments (or populations) with how much variance there is *within* a population.

If the *between* variance is small in comparison to the *within* variance, then there is not strong evidence of a difference between population means.

Variance *between* populations measured by:

$$SSTr = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2.$$

Variance *within* populations measured by

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2.$$

A relationship similar to the one in regression holds:

$$SST = SSTr + SSE,$$

where

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2.$$

SST: total sum of squares; measures *all* the variability in the data, regardless of the source.

$SSTr$: treatment sum of squares; *explained variation*, due to differences between treatments.

SSE: error sum of squares; *unexplained variation*, i.e., variation within a population.

Sums of squares are summarized with an *ANOVA table*, as in regression.

General form of ANOVA table

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Treatments	$k-1$	$SSTr$	$MSTr$	F
Error	$N-k$	SSE	MSE	
Total	$N-1$	SST		

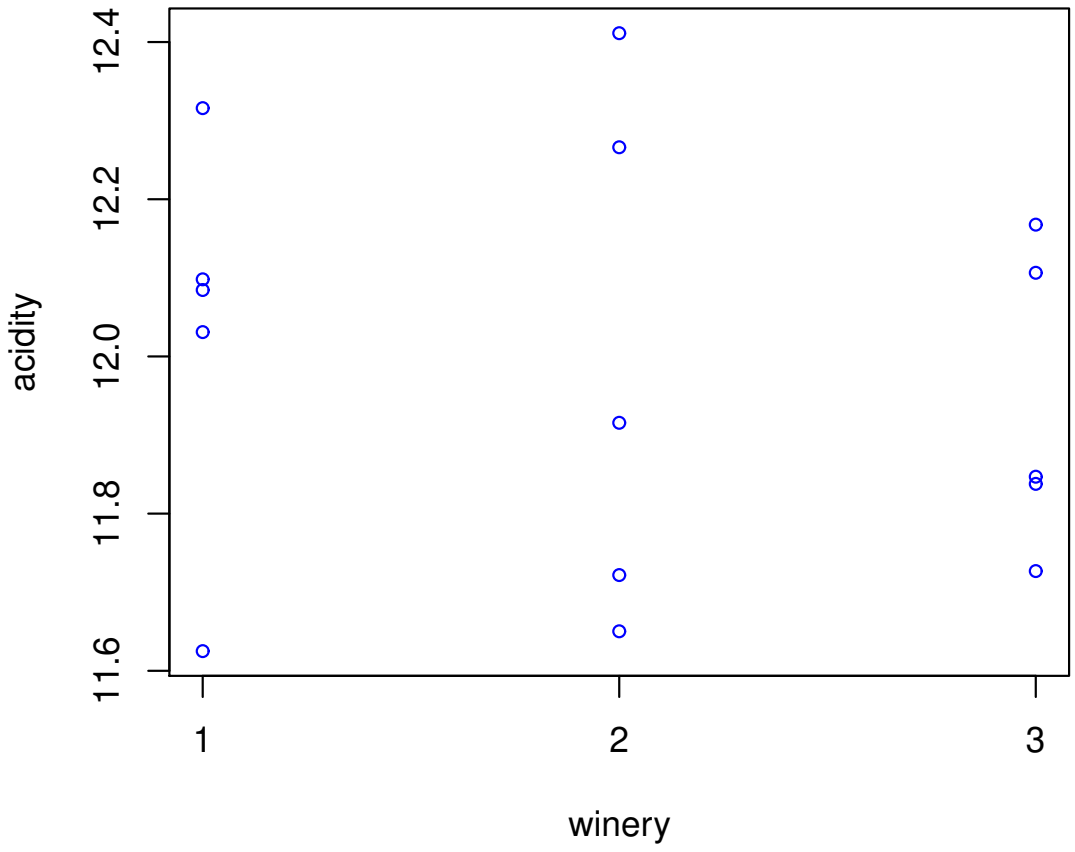
Example 9: *ANOVA for winery data*

It is of interest to compare data from three wineries to find out how wine from the three differs with respect to acidity.

Five bottles of wine are randomly selected from each winery.

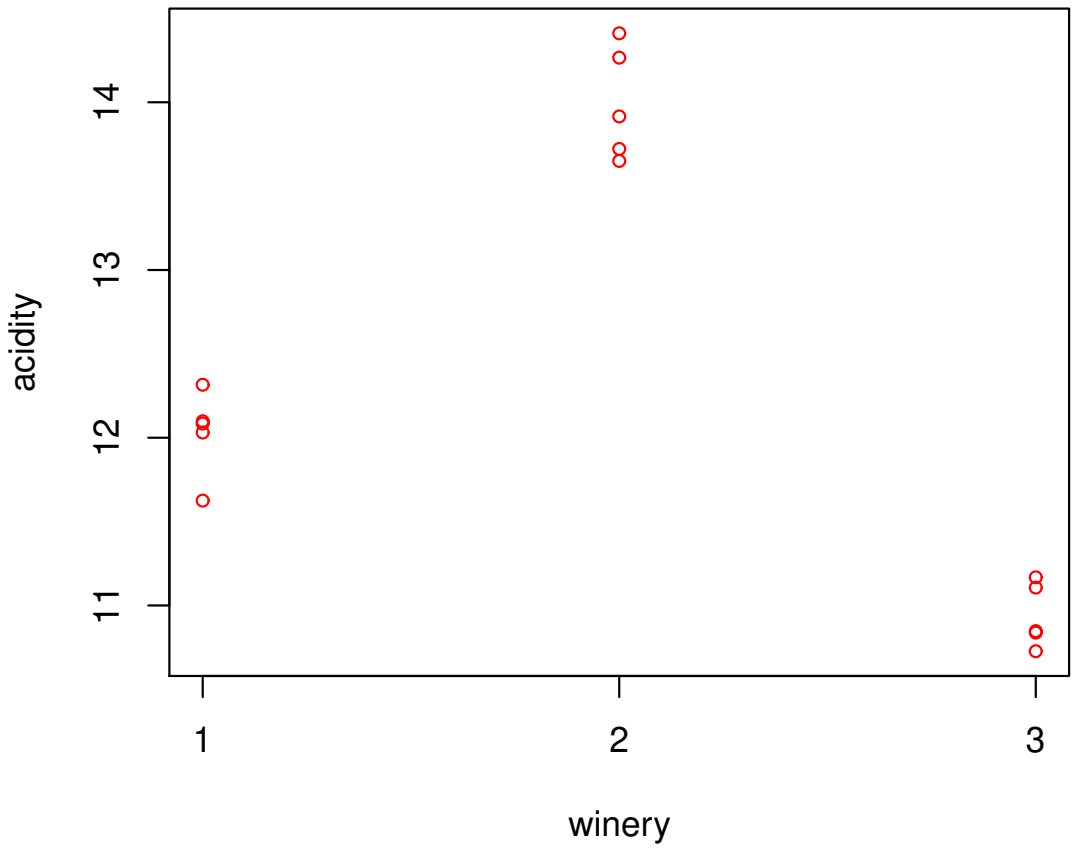
X_{ij} = measure of acidity for jth bottle from winery i

Variation mostly within wineries



Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Wineries	2	0.02224	0.0111	0.158
Error	12	0.84386	0.0703	
Total	14	0.86610		

Substantial variation between wineries



Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Wineries	2	23.9739	11.9870	170.51
Error	12	0.8439	0.0703	
Total	14	24.8178		

Define

$$MSTr = \frac{SSTr}{k-1} \quad \text{and} \quad MSE = \frac{SSE}{N-k}.$$

The F-statistic is:

$$F = \frac{MSTr}{MSE},$$

and the null hypothesis of equal means is rejected at level of significance α when:

$$F \geq F_{k-1,N-k;\alpha}.$$

The P-value is $P(F_{k-1,N-k;\alpha} > F_{obs})$ where F_{obs} is the observed value of the test statistic F.

Motivation for the F-statistic in terms of the model

Recall that

$$\begin{aligned} SSE &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \\ &= \sum_{j=1}^{n_1} (X_{1j} - \bar{X}_1)^2 + \cdots \\ &\quad + \sum_{j=1}^{n_k} (X_{kj} - \bar{X}_k)^2. \end{aligned}$$

The sample variance for the data in group i is

$$S_i^2 = \frac{1}{(n_i - 1)} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2.$$

Therefore

$$SSE = (n_1 - 1)S_1^2 + \cdots + (n_k - 1)S_k^2$$

and

$$\begin{aligned} MSE &= SSE/(N - k) \\ &= \frac{1}{N - k} \sum_{i=1}^{k} (n_i - 1)S_i^2. \end{aligned}$$

1. The last expression says simply that MSE *is a weighted average of the sample variances of the* k *groups*.

2. You may remember from 211 that the sample variance is an unbiased estimator of the population variance.

3. In our one-way ANOVA model, each of the k populations has the same variance, i.e., σ^2.

Together, facts 1, 2 and 3 imply that MSE is an *unbiased estimator* of σ^2.

So, *regardless* of whether or not H_0 is true,

$$E(MSE) = \sigma^2.$$

Now let's see what $MSTr$ estimates.

$$SSTr = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2.$$

According to our model: $X_{ij} = \mu_i + \epsilon_{ij}$, so that:

$$\bar{X}_i = \mu_i + \bar{\epsilon}_i, \quad \text{where } \bar{\epsilon}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \epsilon_{ij},$$

and letting $\bar{\epsilon} = \frac{1}{N}\sum_{i=1}^{k} n_i\bar{\epsilon}_i$, we have:

$$\begin{aligned} \bar{X} &= \frac{1}{N}\sum_{i=1}^{k} n_i\mu_i + \bar{\epsilon} \\ &= \bar{\mu} + \bar{\epsilon}, \quad \text{where } \bar{\mu} = \frac{1}{N}\sum_{i=1}^{k} n_i\mu_i. \end{aligned}$$

So now we have:

$$\bar{X}_i - \bar{X} = (\mu_i - \bar{\mu}) + (\bar{\epsilon}_i - \bar{\epsilon}),$$

and

$$\begin{aligned} (\bar{X}_i - \bar{X})^2 &= (\mu_i - \bar{\mu})^2 + (\bar{\epsilon}_i - \bar{\epsilon})^2 + \\ &\quad 2(\mu_i - \bar{\mu})(\bar{\epsilon}_i - \bar{\epsilon}). \end{aligned}$$

Now we can figure out what the expectation of $SSTr$ is, which will tell us what $MSTr$ is estimating.

$$\begin{aligned} E(SSTr) &= \sum_{i=1}^{k} n_i(\mu_i - \bar{\mu})^2 + E\sum_{i=1}^{k} n_i(\bar{\epsilon}_i - \bar{\epsilon})^2 \\ &= \sum_{i=1}^{k} n_i(\mu_i - \bar{\mu})^2 + (k-1)\sigma^2. \end{aligned}$$

Therefore,

$$E(MSTr) = \sigma^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i(\mu_i - \bar{\mu})^2.$$

This tells us that when all population means are the same (so that $\mu_i = \bar{\mu}$),

$$E(MSTr) = \sigma^2.$$

Otherwise, $MSTr$ is expected to be larger than σ^2.

Recalling the definition of F, we see that *F is expected to be around 1 when all populations have the same mean, and larger than 1 when not all means are the same.*

It would never be sensible to conclude that the population means differ when $F \leq 1$.

Note that all the F critical values in Table A.6 of your text are larger than 1. Unless the degrees of freedom are pretty large, the critical values are quite a bit bigger than 1.

Using R to do one-way ANOVA

- The data file should be such that the first column is all the values of X and the second column indicates which group each X belongs to.

- Suppose your data is in an R dataframe called `Data`, whose first column is `Response` and the second `Factors`. Use the following commands:

```
fit = aov(Response ~ as.factor(Factors),
          data = Data)
anova(fit)
```

Example 10: *Total iron in four iron formations*

On Canvas you'll find a data set giving the total amount of iron in samples from four different iron formations, carbonate, silicate, magnetite and hematite.

We have $k = 4$ and $n_1 = n_2 = n_3 = n_4 = 10$. The following ANOVA table was obtained from R:

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Formations	3	509.12	169.707	10.849
Error	36	563.13	15.643	
Total	39	1072.25		

The P-value for the F-statistic is $3.199\,10^{-5}$.

Since P is much smaller than 0.05 it is reasonable to conclude that the mean amount of iron is not the same for all four formations.

Formation	Mean Iron
Carbonate	26.08
Silicate	24.69
Magnetite	29.95
Hematite	33.84

These means suggest the following rough grouping:

- Carbonate and silicate in one group, and
- magnetite and hematite in another.

It might be that magnetite and hematite differ significantly with respect to mean amount of iron.

We can't answer this question without a more formal procedure for *separating* means.

Multiple comparisons

If we reject

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k,$$

then we'll usually want to know *which* means are different.

Multiple comparisons refers to doing tests of *several* hypotheses, each of the form:

$$H_0 : \mu_i = \mu_j.$$

When testing several hypotheses at once, we should be concerned with the *experimentwise error-rate (EWER)*, which is the probability of rejecting *at least one* of the null hypotheses when in fact they are *all* true.

Also known as *familywise error rate (FWER).*

Example 11: *Testing several hypotheses independently*

Suppose we want to test m null hypotheses, call them $H_0^1, \ldots, H_0^m$. We have m independent sets of data with which to test the hypotheses.

Practical situation where this arises: m different labs do independent experiments, all aimed at testing the same hypothesis.

Suppose a size α test of each hypothesis is performed. This means that, for each i,

$$P(\text{rejecting } H_0^i | H_0^i \text{ is true}) = \alpha.$$

Now let

$\alpha_E =$ experimentwise error rate($EWER$).

By definition:

$$\begin{aligned} \alpha_E &= P(\text{rejecting } \underline{\textit{at least}} \\ &\qquad \underline{\textit{one}}\ H_0^i \mid \underline{\textit{all}}\ H_0^i \text{ are true}). \end{aligned}$$

Using the law of the complement,

$$\begin{aligned} \alpha_E &= 1 - P(\text{fail to reject } \underline{\textit{each}}\ H_0^i | \ldots) \\ &= 1 - (1-\alpha)^m. \end{aligned}$$

Why is the last step true?

When $\alpha = 0.05$, we have the following experimentwise error rates (EWERs):

m	α_E
1	0.0500
2	0.0975
3	0.1426
4	0.1855
5	0.2262
10	0.4013
15	0.5367
20	0.6415
30	0.7854
40	0.8715
50	0.9231

A method for doing multiple comparisons is *Tukey's procedure (Section 10.3.2).*

This method allows the experimentwise error rate to be set at any desired level α.

Steps of Tukey's procedure:

1. Choose the error rate (EWER). Call it α.

2. Find critical value $Q_{\alpha,k,N-k}$ from Table A.7.

3. For all (i, j), compute the following confidence interval for $\mu_i - \mu_j$:

$$\bar{X}_i - \bar{X}_j \pm Q_{\alpha,k,N-k}\sqrt{\frac{MSE}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

4. For each pair (i, j) such that the interval in **3** does *not* contain 0, reject $H_0 : \mu_i = \mu_j$.

This approach controls the EWER at level α.

Example 12: Use of Tukey's procedure

The data for this example, which may be found at the website, are yields (in lbs.) for five varieties of orange trees.

Yields from seven different trees are obtained for each variety.

So, we have $k = 5$ and $n = 7$. Take $\alpha = 0.05$.

In R suppose I define the response to be `pounds` and the factor variable to be `varieties`. Then we may obtain 95% confidence intervals for all $\mu_i - \mu_j$ using Tukey's procedure with the command

`TukeyHSD(aov(pounds ~ varieties))`.

A nice plot of the results is gotten with

`plot(TukeyHSD(aov(pounds ~ varieties)))`.

Our conclusion is that *average yield for variety C is significantly higher than average yield for variety A, but no other pair of varieties has significantly different average yields.*