STA 111: Probability & Statistical Inference

Lecture Twenty – Analysis of Variance
D.S. Sections 11.6, 11.7 & 11.8

Instructor: Olanrewaju Michael Akande

Department of Statistical Science, Duke University

Outline

- Questions from Last Lecture
- One-Way ANOVA
- Illustration
- RCBD
- Another Illustration

Motivation

Experiments are often expensive and/or dangerous. One wants to use good techniques that minimize costs and risks, while still learning efficiently about the treatment(s) of interest.

Experimental design was invented by Sir Ronald Fisher, in the context of agricultural experiments at Rothamstead Experimental Station (RES). The RES scientists were involved in improving crop yield through the study of better strains, fertilizers, crop rotation, irrigation, and procedures for tilling, weeding, and harvesting.

When Fisher joined RES, it had been operating for about 80 years. RES employees knew that their fields had fertility gradients, different exposures to sunlight, and that yield could be affected by crops and practices used in previous seasons.

Motivation

It was not sufficient to pick one field at random and use, say, fertilizer A, and another at random and use fertilizer B, then compare the yields. The observed difference could be due to sunlight exposure or previous crops.

In principle one could pick many fields at random for fertilizer A, and many fields at random for fertilizer B, and then compare the results.

But it is even better if you can pick the fields at random in such a way that the choice controls for possible confounding factors; e.g., it would be good if equal numbers of fertilizer A fields and fertilizer B fields were in bright sunshine (or shaded), or had good drainage (or not), and so forth.

This kind of balance ensures that one can learn a lot from a small number of observations. And this is important in agricultural experiments where each dataset takes about a year to collect and farm acreage is a resource constraint.

ANOVA

The Analysis of Variance (ANOVA) is a general term for a statistical strategy for analyzing data collected from designed experiments such as those at RES.

There is a ladder of complexity:

- one-way ANOVA,
- two-way ANOVA without interaction (or Randomized Complete Block Designs),
- two-way ANOVA with interaction
- higher-order designs
- fractional factorials.

We will discuss the first two.

Definitions

Some definitions:

An **experimental unit** is the smallest amount of material to which a single treatment can be applied. In the case of RES, this is usually a plot (a small field), but in other applications it might be a person or a mouse or a school.

A **treatment** is a factor we are interested in. A placebo is a special case of a treatment. In general, the manipulation is very consistent; e.g., in Rothamsted, a treatment might be adding 20 pounds of fertilizer A per square acre.

An **observation** is a measurement taken upon an experimental unit. Typically, an observation reflects some measurement error, the effects of the treatment applied to that experimental unit, and the results of unmeasured variables (such as fertility gradients, drainage, and unknown factors).

Notation

Some notation:

- I is the number of treatments.
- X_{ij} is the measurement on the jth unit receiving treatment i.
- n_i is the number of experimental units that received treatment i.
- $n_.$ is the total number of observations.
- $X_{i.}$ is the sum of all measurements for units receiving treatment i.
- $\bar{X}_{i.}$ is the average of all measurements for units receiving treatment i, or $\frac{1}{n_i}\sum_{j=i}^{n_i} X_{ij}$.
- $\bar{X}_{..}$ is the average of all measurements.

Note that whenever we sum over a subscript, we replace the subscript by a dot. Similarly, when we average over a subscript, we indicate that by a bar and the average is taken with respect to the dotted subscript.

Assumptions

Some assumptions:
The model for one-way ANOVA is:

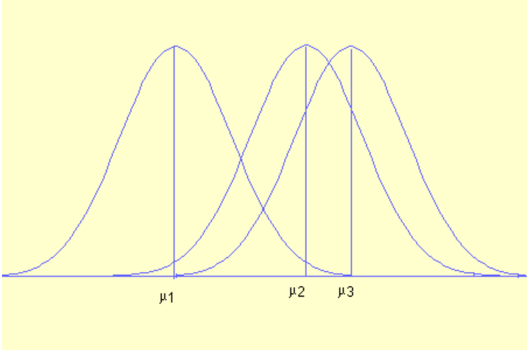
$$\begin{aligned} X_{ij} &= \mu_i + \epsilon_{ij} \\ &= \mu + \tau_i + \epsilon_{ij} \end{aligned}$$

where μ is the overall mean of all experimental units, τ_i is the effect of treatment i, and the ϵ_{ij} are random errors that are:

- normally distributed with mean zero and unknown standard deviation σ_ϵ
- independent of the values for all other errors.

Note that this has some formal similarity to the regression model. It turns out that ANOVA is a special case of regression.

Testing



The image above corresponds to a one-way ANOVA with three treatments. Observations are taken from each of the three distributions, and we wish to decide from those observations whether there is evidence that the means of the three groups are different.

Testing

The hypothesis test in the one-way ANOVA is always the same:
H_0: The means of all treatments are equal (or H_0: $\mu_1 = \cdots = \mu_I$; or H_0: $\tau_1 = \cdots = \tau_I = 0$)
versus
H_A: At least one of the treatments has a different mean
The strategy in making an ANOVA test is to partition the total variation in the data into components attributable to different effects.

In the case of one-way ANOVA, we divide the total sum of squares (SS_{tot}) into the part attributable to differences between the treatment means (SS_{trt}) and the part attributable to differences within treatment groups, or the sum of squares due to pure error (SS_{err}) .

Testing

The test statistic for a one-way ANOVA test looks at the standardized ratio of the between-group sum of squares to the within-group sum of squares. If this is large, then it suggest that some of the treatments are not equal.

Formally, the test statistic is

$$ts = \frac{\mathbf{SS_{trt}}/(I-1)}{\mathbf{SS_{err}}/(n_{.}-I)}.$$

This is compared to an F distribution with $I-1$ degrees of freedom in the numerator and $n_{.}-I$ degrees of freedom in the denominator. Note that the total degrees of freedom is $n_{.}-1$; we have lost one degree of freedom due to estimating μ, the overall mean of the combined populations.
To calculate this test statistic, use the following definitional and computational formula.

Testing

$$
\begin{aligned}
\textbf{SS}_{\textbf{tot}} &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{..})^2\\
&= \left(\sum_{i=1}^{I}\sum_{j=1}^{n_i}X_{ij}^2\right)-\frac{X_{..}^2}{n_.}\\
\textbf{SS}_{\textbf{trt}} &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{X}_{i.}-\bar{X}_{..})^2=\sum_{i=1}^{I}n_i(\bar{X}_{i.}-\bar{X}_{..})^2\\
&= \left(\sum_{i=1}^{I}\frac{X_{i.}^2}{n_i}\right)-\frac{X_{..}^2}{n_.}\\
\textbf{SS}_{\textbf{err}} &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i.})^2\\
&= \left(\sum_{i=1}^{I}\sum_{j=1}^{n_i}X_{ij}^2\right)-\sum_{i=1}^{I}\frac{X_{i.}^2}{n_i}
\end{aligned}
$$

Testing

Note that $\text{SS}_{\text{tot}} = \text{SS}_{\text{trt}} + \text{SS}_{\text{err}}$. This is useful because it simplifies calculation—if you know two the sum of square terms, you can find the third by subtraction.

Once you have calculated the test statistic, you refer it to an F distribution. For example, suppose we had a total of 10 observations on three different groups. Then $I = 3$ and $n_{\cdot} = 10$ so the F distribution we should use had 2 df in the numerator and 7 in the denominator.

From the F-table, we should reject the null hypothesis that all treatment means are equal at the 0.05 level if the ts is greater than 4.74.

Note that all ANOVA tests are one-sided just like we had for Chi-squared tests – we reject if and only if we get large values of the test statistic. Also note that if we had only two groups, then this would informationally equivalent to a two-sample t-test, which is why an F with 1 df in the numerator and k df in the denominator is the square of a t with k df.

One-way ANOVA Table

To simplify the organization of the calculations in ANOVA, it is customary to write things in an ANOVA table.

Source	df	SS	MS	F
treatment	$I-1$	SS_{trt}	$\text{SS}_{\text{trt}}/(I-1)$	$\text{MS}_{\text{trt}}/\text{MS}_{\text{err}}$
error	$n_{\cdot}-I$	SS_{err}	$\text{SS}_{\text{err}}/(n_{\cdot}-I)$	
total	$n_{\cdot}-1$	SS_{tot}		

The MS column contains the **Mean Squares**, which are the average sum of squares attributable to each component in the partition. The F column contains the test statistic.

The table is not especially useful for simple one-way ANOVA, but it is very helpful as the designs get more complex.

Examples

Example 1: Suppose you have three different feeds that may affect the size of eggs that chickens lay. You randomly assign 10 chickens to one of the three feeds and record the size of the eggs (maximum length, in centimeters) that the chickens lay the following week.

The data are:
Feed A: 7 cm, 6 cm, 6 cm
Feed B: 9 cm, 8 cm, 7 cm
Feed C: 9 cm, 10 cm, 10 cm, 11 cm

Examples

The null hypothesis is that all the chicken feeds have the same effect on the length of the major axis. The alternative is that the feed has some causal effect.

We use the computational forms for the sum of squares calculation. Thus:

$$
\begin{aligned}
\sum_{i=1}^{I}\sum_{j=1}^{n_i} X_{ij}^2 &= 717 \\
\frac{X_{..}^2}{n_.} &= 688.9 \\
\sum_{i=1}^{I}\frac{X_{i.}^2}{n_i} &= 712.33
\end{aligned}
$$

Therefore $\text{SS}_{\text{tot}} = 717\text{-}688.9 = 28.1$, $\text{SS}_{\text{trt}} = 712.33 - 688.9 = 23.43$, and $\text{SS}_{\text{err}} = 4.67$.

Examples

We plug this into the ANOVA table:

Source	df	SS	MS	F
feed	2	23.43	11.715	17.56
error	7	4.67	.667	
total	9	28.1		

Our test statistic 17.56 is larger than $F_{2,7;.05} = 4.74$. We reject the null—there is strong evidence that the feed affects egg size.

Examples

Example 2 (D.S. Section 11.6 Exercise, Question 4 – to be done in class): Specimens of milk from a number of dairies in three different districts were analyzed, and the concentration of the radioactive isotope strontium-90 was measured in each specimen. Suppose that specimens were obtained from four dairies in the first district, from six dairies in the second district, and from three dairies in the third district, and that the results measured in picocuries per liter were as follows:

District 1: 6.4, 5.8, 6.5, 7.7
District 2: 7.1, 9.9, 11.2, 10.5, 6.5, 8.8
District 3: 9.5, 9.0, 12.1

Test the hypothesis that the three districts have identical concentrations of strontium-90.

Multiple Comparisons

If one has completed a one-way ANOVA and has rejected the null hypothesis of no treatment differences, the immediate question is which treatment (or treatments) is best?

In the formal framework described so far, that question cannot be answered without eroding the alpha level of the test, so that the overall error rate is larger than the nominal value (say 0.05) used in the original ANOVA or RCBD.

To address this problem, statisticians have developed procedures for **multiple comparisons**. Some of these get quite complicated. We won't review them here.

Two-way ANOVA without Interaction

The **Randomized Complete Block Design** is also known as the two-way ANOVA without interaction. A key assumption in the analysis is that the effect of each level of the treatment factor is the same for each level of the blocking factor.

That assumption would be violated if, say, a particular fertilizer worked well for one stain but poorly for another; or if one cancer therapy were better for lung cancer but a different therapy were better for stomach cancer.

In RCBD, there is one observation for each combination of levels of the treatment and block factors.

Notation

RCBD notation:

- I is the number of treatments; J is the number of blocks
- X_{ij} is the measurement on the unit in block j that received treatment i.
- $n_{..} = I * J$ is the total number of experimental units.
- $n_{i.} = J$ and $n_{.j} = I$, the number of obersvations for a given treatment level or block level, respectively.
- $X_{i.}$ is the sum of all measurements for units receiving treatment i, and $X_{.j}$ is the sum of all measurements for units in the j block.
- $\bar{X}_{i.}$ is the average of all measurements for units receiving treatment i, or $\frac{1}{J}\sum_{j=1}^{J} X_{ij}$.
- $\bar{X}_{.j}$ is the average of all measurements for units in the jth block, or $\frac{1}{I}\sum_{i=1}^{I} X_{ij}$.
- $\bar{X}_{..}$ is the average of all measurements.

We continue to use the dot and bar conventions.

Assumptions

Some assumptions:
The model for an RCBD (or one-way ANOVA without interactions) is:

$$X_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}$$

where μ is the overall mean of all experimental units, τ_i is the effect of treatment i, β_j is the effect of block j, and the ϵ_{ij} are random errors that are:

-
- normally distributed with mean zero and unknown standard deviation σ
- independent of the values for all other errors.

Note how this generalizes the one-way ANOVA model:

$$X_{ij} = \mu + \tau_i + \epsilon_{ij}.$$

Testing

There are two hypothesis tests in an RCBD, and they are always the same:
H_0: The means of all treatments are equal or H_0: $\tau_1 = \cdots = \tau_I = 0$
versus
H_A: At least one of the treatments has a different mean

and

H_0: The means of all blocks are equal or H_0: $\beta_1 = \cdots = \beta_J = 0$.
versus
H_A: At least one of the blocks has a different mean

In the case of an RCBD, we divide the total sum of squares (SS_{tot}) into the part attributable to differences between the treatment means (SS_{trt}), the part attributable to difference between the blocks (SS_{blk}) and the part attributable to differences within block-treatment groups, or the sum of squares due to pure error (SS_{err}).

Testing (Cont'd)

The test statistics for an RCBD look at the standardized ratio of the between-group sum of squares to the within-group sum of squares for the treatment and block effects separately. If one or both are large, then it suggest that some of the treatment or block effects are not equal.

Formally, the test statistics are:

$$ts_{\textbf{trt}} = \frac{\textbf{SS}_{\textbf{trt}}/(I-1)}{\textbf{SS}_{\textbf{err}}/(I-1)(J-1)}$$

and

$$ts_{\textbf{blk}} = \frac{\textbf{SS}_{\textbf{blk}}/(J-1)}{\textbf{SS}_{\textbf{err}}/(I-1)(J-1)}.$$

Testing (Cont'd)

These two test statisics are compared to an F distribution with $I-1$ or $J-1$ (respectively) degrees of freedom in the numerator and $IJ-I-J+1$ degrees of freedom in the denominator.

The numerator degrees of freedom for the test of treatments is $I-1$. That is because we make $I-1$ estimates to get it; one estimate for each of the I different groups, but the last one is not needed since the sum of the treatment effects is forced to add up to zero (since we have already found the overall mean, which cost us one 1 degree of freedom).

Similarly, the numerator degrees of freedom for the block term is $J-1$.

Since that the total degrees of freedom is $IJ-1$, then the degrees of freedom for the error term in the denominator must be $(IJ-1)-(I-1)-(J-1)=IJ-I-J+1=(I-1)(J-1)$. As always, we have lost one degree of freedom due to estimating μ, the overall mean of the combined populations.

Testing (Cont'd)

To understand this test statistic, consider the following definitional formulae:

$$
\begin{aligned}
\mathbf{SS_{tot}} &= \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{..})^2 \\
\mathbf{SS_{trt}} &= \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i=1}^{I} J(\bar{X}_{i.} - \bar{X}_{..})^2 \\
\mathbf{SS_{blk}} &= \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{X}_{.j} - \bar{X}_{..})^2 = \sum_{j=1}^{J} I(\bar{X}_{.j} - \bar{X}_{..})^2 \\
\mathbf{SS_{err}} &= \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2
\end{aligned}
$$

Testing (Cont'd)

In practice, if we aren't using a computer, then it is easier, faster, and more numerically stable to use the following computational formulae:

$$\begin{aligned}
\mathbf{SS_{tot}} &= \left(\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}^2\right) - \frac{X_{\cdot\cdot}^2}{IJ}\\
\mathbf{SS_{blk}} &= \sum_{j=1}^{J}\frac{X_{\cdot j}^2}{I} - \frac{X_{\cdot\cdot}^2}{IJ}\\
\mathbf{SS_{trt}} &= \sum_{i=1}^{I}\frac{X_{i\cdot}^2}{J} - \frac{X_{\cdot\cdot}^2}{IJ}\\
\mathbf{SS_{err}} &= \left(\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}^2\right) - \sum_{i=1}^{I}\frac{X_{i\cdot}^2}{J} - \sum_{j=1}^{J}\frac{X_{\cdot j}^2}{I} + \frac{X_{\cdot\cdot}^2}{IJ}
\end{aligned}$$

Testing (Cont'd)

Note that $\text{SS}_{\text{tot}} = \text{SS}_{\text{trt}} + \text{SS}_{\text{blk}} + \text{SS}_{\text{err}}$. This is useful because it simplifies calculation—if you know three of the sum of square terms, you can find the fourth by subtraction.

Once you have calculated the test statistics, you refer them to F distributions. For the test of equal treatment effects, the numerator df is $I-1$. For the test of equal block effects, the numerator df is $J-1$. Both tests have the same denominator df, $(I-1)(J-1)$.

Note that both RCBD tests are one-sided—we reject if and only if we get large values of the test statistic. Also note that if we had only two treatments, then this would be informationally equivalent to a paired difference two-sample t-test and use an F with 1 df in the numerator and $J-1$ df in the denominator.

Two-way ANOVA Table

To simplify the organization of the calculations in an RCBD, it is customary to write things in a table.

Source	df	SS	MS	F
treatment	$I-1$	SS_{trt}	$\text{SS}_{\text{trt}}/(I-1)$	$\text{MS}_{\text{trt}}/\text{MS}_{\text{err}}$
block	$J-1$	SS_{blk}	$\text{SS}_{\text{blk}}/(J-1)$	$\text{MS}_{\text{blk}}/\text{MS}_{\text{err}}$
error	$(I-1)(J-1)$	SS_{err}	$\text{SS}_{\text{err}}/(I-1)(J-1)$	
total	$IJ-1$	SS_{tot}		

The MS column contains the **Mean Squares**, which are the average sum of squares attributable to each component in the partition. The F column contains the test statistic.

What happens if the block effect is not significant?

Two-way ANOVA Example

Example 3: Suppose you are manufacturing concrete cylinders for, say, bridge supports. There are three ways of drying green concrete (say A, B, and C), and you want to find the one that gives you the best compressive strength. The concrete is mixed in batches that are large enough to produce exactly three cylinders, and your production engineer believes that there is substantial variation in the quality of the concrete from batch to batch.

You have data from $J = 5$ batches on each of the $I = 3$ drying processes. Your measurements are the compressive strength of the cylinder in a desctructive test. (So there is an economic incentive to learn as much as you can from a well-designed experiment.)

Two-way ANOVA Example

The data are:

	Batch					
Treatment	1	2	3	4	5	Trt Sum
A	52	47	44	51	42	236
B	60	55	49	52	43	259
C	56	48	45	44	38	231
Batch Mean	168	150	138	147	123	726

The primary null hypothesis is that all three drying techniques are equivalent, in terms of conferring compressive strength. The secondary null is that the batches are equivalent (but if they are, then we have wasted power by controlling for an effect that is small or non-existent).

Two-way ANOVA Example

We use the computational forms for the sum of squares calculation. Thus:

$$
\begin{aligned}
\mathbf{SS_{tot}} &= \left(\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^2\right) - \frac{X_{..}^2}{IJ} \\
&= 499.6 \\
\mathbf{SS_{blk}} &= \sum_{j=1}^{J} \frac{X_{.j}^2}{I} - \frac{X_{..}^2}{IJ} \\
&= 363.6 \\
\mathbf{SS_{trt}} &= \sum_{i=1}^{I} \frac{X_{i.}^2}{J} - \frac{X_{..}^2}{IJ} \\
&= 89.2 \\
\mathbf{SS_{err}} &= \left(\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^2\right) - \sum_{i=1}^{I} \frac{X_{i.}^2}{J} - \sum_{j=1}^{J} \frac{X_{.j}^2}{I} + \frac{X_{..}^2}{IJ} \\
&= 46.8
\end{aligned}
$$

Two-way ANOVA Example

We plug this into the ANOVA table:

Source	df	SS	MS	F
drying	2	89.2	44.6	7.62
batch	4	363.6	90.9	15.54
error	8	46.8	5.85	
total	14	499.6		

Our test statistics are 7.62 and 15.54. The test of the first uses an F with 2 df in the numerator, 8 in the denominator, so the 0.05 critical value is 4.46. We reject.

The secondary test has 4 df in the numerator, 8 in the denominator, and the 0.05 critical value is 3.84. The blocking was a good idea.

Two-way ANOVA Example

Suppose we had not blocked for batch. Then the data would be:

Treatment		Trt Sum
A	52, 47, 44, 51, 42	236
B	60, 55, 49, 52, 43	259
C	56, 48, 45, 44, 38	231

This is the same as before except now we ignore which batch the observation came from.

Two-way ANOVA Example

The one-way ANOVA table for this is:

Source	df	SS	MS	F
drying	2	89.2	44.6	1.30
error	$4 + 8$	46.8+363.6	34.2	
total	14	499.6		

Note that this gives a test statistic of 1.30, which is referred to an F-distribution with 2 df in the numerator, 12 in the denominator. The .05 critical value is 3.89. We fail to reject the null.

Using blocks gave us a more powerful test. Remember Simpson's Paradox?