

AnSc 5403
Biometry

Lecture Notes 19

I. Introduction to Analysis of Variance (AOV)

A. Thus far, we have been concerned only with confidence intervals and tests of hypotheses on a single population parameter or the differences in populations parameters between two groups

1. What happens when we have more than two groups?

B. Frequently, we want to test the null hypothesis that there is –

1. The method typically used to test a hypothesis of this type is the analysis of variance

C. The basis of the analysis of variance

1. In its simplest form, AOV asks –

D. Let's consider an example

Trt 1	Trt 2	Trt 3	Trt 4
22	36	24	25
30	30	21	30
43	17	35	32
32	40	16	33
<hr/>			
$\bar{x} = 31.75$	$\bar{x} = 30.75$	$\bar{x} = 24.00$	$\bar{x} = 30.00$
$s^2 = 74.91$	$s^2 = 100.91$	$s^2 = 64.66$	$s^2 = 12.66$

1. Now, assuming that the within-treatment variability is the same in all four treatment groups (e.g., assuming equality of variance), we can estimate the population variance in one of two ways:

a. Estimate 1

$$(a) s^2_{\text{pooled}} = (3 \times 74.91 + 3 \times 100.91 + 3 \times 64.66 + 3 \times 12.66)/(3 + 3 + 3 + 3)$$

(b) s^2 pooled =

b. Estimate 2 (this one involves an assumption)

(a) If we assume that there is no difference in the true means among the four treatment groups, then we can multiply the variability in the observed means by four (the number of treatments) and we should get the same answer as s^2 pooled

(i) Reasoning:

1. We know that $\sigma^2_{\bar{x}} = \sigma^2/4$

2. So, by simple math, $\sigma^2 = 4 \times \sigma^2_{\bar{x}}$

(ii) We can calculate $s^2_{\bar{x}}$ just like we would for any estimate of variance, but just using the four treatment means

1. $s^2_{\bar{x}} = \{(31.75)^2 + (30.75)^2 + (24.00)^2 + (30.00)^2 - [(116.50)^2/4]\}/3$

2. $s^2_{\bar{x}} = 12.18$ and $4 \times s^2_{\bar{x}} =$

c. So, now we have s^2 pooled, the variability we expect –

d. And we also have $s^2_{\bar{x}}$ - the variability we expect

e. In this case, these two estimates are fairly similar

2. Now, let's take our data and see what happens when we create some differences among the four treatment means

a. To each observation, add 10 to Trt 2, 20 to Trt 3, and 30 to Trt 4

Trt 1	Trt 2	Trt 3	Trt 4
22	46	44	55
30	40	41	60
43	27	55	62
32	50	36	63

$\bar{x} = 31.75$	$\bar{x} = 40.75$	$\bar{x} = 44.00$	$\bar{x} = 60.00$
$s^2 = 74.91$	$s^2 = 100.91$	$s^2 = 64.66$	$s^2 = 12.66$

b. Remember that if we add a constant, the sample variance does not change, which is why the s^2 values remain the same

3. So, our s^2 pooled is the same (63.29), but what about our $s^2_{\bar{x}}$?

a. $s^2_{\bar{x}} = \{(31.75)^2 + (40.75)^2 + (44.00)^2 + (60.00)^2 - [(176.50)^2/4]\}/3$

b. $s^2_{\bar{x}} =$

c. So, $4 \times s^2_{\bar{x}} =$

(a) This value is approximately – eight to nine times the s^2 pooled value of 63.29

4. So, it seems as if the ratio of these two estimates (s^2 pooled and $4 \times s^2_{\bar{x}}$) gives us an indication of whether there are differences among the four treatment means

a. This is the basic idea (logic) behind the AOV

b. We will use some shortcut methods to make the calculations easier, but this is the basic approach

(a) Before we take a look at those shortcut methods, let's consider one other way of looking at the principles behind the AOV

E. The linear model concept

1. Consider the example we had above, which is often referred to as a –
 - a. This design would be one in which we have several experimental units to which we randomly apply treatments
 - (a) By definition, the experimental unit is – *the smallest unit or entity to which a given treatment was applied*
 - (i) In our case, to a total of 16 experimental units, we applied four treatments (Groups 1 through 4) at random and then made measurements
 - b. We should be able to describe any observation in our group of 16 as a function of the group mean and the deviation of that observation from the group mean
 - (a) We could write this as: $X_{ij} = \mu_i + \varepsilon_{ij}$
 - (i) Where X_{ij} = the j-th observation from the i-th population
 - (ii) μ_i = the mean value of the i-th population
 - (iii) $\varepsilon_{ij} = X_{ij} - \mu_i$ = the deviation of the i,j-th observation from the mean
 - c. In our case, however, we applied four treatments to the groups, and we are interested in determining the effect of those treatments
 - (a) Based on our previous example, it should be clear that what we want to know is how much each treatment mean deviates from the overall mean
 - (b) So, if we define the overall mean as μ and the treatments means as μ_i , then each observation can be defined as:
 - (i) $X_{ij} = \mu + (\mu_i - \mu) + \varepsilon_{ij}$
 - (ii) And if we define $(\mu_i - \mu)$ as τ_i , we can rewrite this model as:
 1. $X_{ij} = \mu + \tau_i + \varepsilon_{ij}$
 - (c) The τ_i component of this model might be a one of two types of effects:
 - (i) A fixed effect – all treatments occur every time the experiment is conducted and we want to make inference to that set of treatments
 1. Example – we want to compare four varieties of corn and make inference about the effects of those specific varieties

(ii) A random effect – different sets of treatments occur in different experiments and we want to make inference to some larger set from which the experimental treatments were selected

1. An engineer might compare four temperatures in a given experiment, but wants to make inference to all temperatures in the range covered

(iii)NOTE – Whether the treatment effect is fixed or random affects the way we do testing in the AOV, which is a bit beyond our scope, and not particularly relevant to the completely random design

1. However, this difference between fixed and random effects and the fact that it affects how testing is done will be important as one's statistical horizons expand!

2. So, in the context of our previous example, let's assume that the treatment effects and fixed

- a. We calculated two estimates of the variance – one based on s^2 pooled and one based on $4 \times s^2_{\bar{x}}$
- b. It turns out that our s^2 pooled is an estimate of – *the variance associated with ϵ_{ij} in our linear model*
- c. And $4 \times s^2_{\bar{x}} =$ an estimate of –

(a) If we have r observations in each treatment group, the τ_i component in our model is actually calculated as $-\sum r_i \tau_i^2 / (k - 1)$, where k is the number of treatments

3. So, if we take the ratio of $\tau_i + \sigma^2_{\epsilon_{ij}}$ and $\sigma^2_{\epsilon_{ij}}$, it should tell us something about the magnitude of the τ_i

- a. Just in case you were wondering, in a random-effects model, the $4 \times s^2_{\bar{x}}$ component would be an estimate of the $r_0 \sigma^2_{\tau_i} + \sigma^2_{\epsilon_{ij}}$, where r_0 is a function of the total number of observations and the number of treatment groups, given by the formula – $r_0 = (n - \sum r_i^2 / n) / (k - 1)$

4. Remember when we take the ratio of variances, the distribution of interest to determine probability values is the –

F. The linear model concept leads us to the ASSUMPTIONS of the AOV

1. For inference to be valid using the AOV procedures, the following assumptions must hold (from Elementary Applied Statistics, 1972 – M. Lentner, Bogden and Quigley, Inc., Publishers, Tarrytown-on-Hudson, NY):

- a. The linear model is appropriate
- b. The experimental units are –
- c. μ is a fixed constant that is common to all observations
- d. The ε_{ij} are –
- e. For a fixed-effects model, the τ_i are fixed constants such that $-\sum r_i \tau_i = 0$, where r_i is the number of observations per treatment mean
 - (a) In the random-effects case, the τ_i are normally and independently distributed with $\mu = 0$ and variance σ_τ^2 , and the τ_i are distributed independently of the ε_{ij} 's