Analysis of Variance (ANOVA)

DR.ALAA MOHAMMED

3RD LEVEL, 2ND SEMESTER, BIOMEDICAL INSTRUMENTATION AND BIOMECHANICS BRANCHES, BIOMEDICAL ENG. DEPARTMENT

Analysis of Variance (ANOVA)

Analysis of Variance (**ANOVA**) is a hypothesis-testing technique used to test the equality of two or more population (or treatment) means by examining the variances of samples that are taken.

ANOVA allows one to determine whether the differences between the samples are simply due to random error (sampling errors) or whether there are systematic treatment effects that causes the mean in one group to differ from the mean in another.

ANOVA is based on comparing the variance (or variation) *between* the data samples to variation *within* each particular sample.

If the between variation is much larger than the within variation, the means of different samples will not be equal.

Assumptions of ANOVA

I. All populations involved follow a normal distribution.

II. All populations have the same variance (or standard deviation).

III. The samples are randomly selected and independent of one another.

Basic ANOVA concepts

❖**The Setting**

Generally, we are considering a quantitative response variable as it relates to one or more explanatory variables, usually categorical. Questions which t this setting:

- Which academic department in the sciences gives out the lowest average grades? (Explanatory variable: *department*; Response variable: *student GPA's for individual courses*)
- Which kind of promotional campaign leads to greatest store income at Christmas time? (Explanatory variable: *promotion type*; Response variable: *daily store income*)
- How do the type of career and marital status of a person relate to the total cost in annual claims she/he is likely to make on her health insurance. (Explanatory variables: *career* and *marital status*; Response variable: *health insurance payouts*)

Basic ANOVA concepts (cont.)

❖**Hypotheses of ANOVA**

These are always the same.

➢*Ho*: The (population) means of all groups under consideration are equal.

➢*Ha*: The (pop.) means are not all equal.

(Note: This is different than saying .they are all unequal .!)

One-Way ANOVA

One-way ANOVA examines equality of population means for a quantitative out-come and a single categorical explanatory variable with any number of levels.

The term **one-way**, also called one-factor, indicates that there is a single explanatory variable (“treatment") with two or more levels, and only one level of treatment is applied at any time for a given subject.

We use the term two-way or two-factor ANOVA, when the levels of two different explanatory variables are being assigned, and each subject is assigned to one level of each factor.

It is worth noting that the situation for which we can choose between **one-way ANOVA** and an independent samples **t-test** is when the explanatory variable has exactly two levels. In that case we always come to the same conclusions regardless of which method we use.

How one-way ANOVA works

The model and statistical hypothesis

One-way ANOVA is appropriate when the following model holds.

- We have a single “treatment" with, say, k levels.
- “Treatment" may be interpreted in the loosest possible sense as any categorical explanatory variable.
- The population variances for the outcome for each of the k groups defined by the levels of the explanatory variable all have the same value, usually called σ^2, with no restriction other than that $\sigma^2 > 0$.
- For treatment i, the distribution of the outcome is assumed to follow a Normal distribution with mean μ_i and variance σ^2, often written $N(\mu_i, \sigma^2)$.

HOW ONE-WAY ANOVA WORKS

Technically, the sample group means are unbiased estimators of the population group means when treatment is randomly assigned.

The meaning of unbiased here is that the true mean of the sampling distribution of any group sample mean equals the corresponding population mean.

Further, under the Normality, independence and equal variance assumptions it is

true that the sampling distribution of $\overline{Y_i}$ is $N(\mu_i, \sigma^2/n_i)$, exactly.

The statistical model for which one-way ANOVA is appropriate is that the (quantitative) outcomes for each group are normally distributed with a common variance σ^2.The errors (deviations of individual outcomes from the population group means) are assumed to be independent. The model places no restrictions on the population group means.

The term assumption in statistics refers to any specific part of a statistical model.

For one-way ANOVA, the assumptions are normality, equal variance, and independence of errors.

Correct assignment of individuals to groups is sometimes considered to be an implicit assumption.

The null hypothesis

The **null hypothesis** is a point hypothesis stating that “nothing interesting is happening.“

For one-way ANOVA, we use H_0: $\mu_1 = \ldots = \mu_k$, which states that all of the population means are equal, without restricting what the common value is.

This null hypothesis is called the “overall" null hypothesis and is the hypothesis tested by ANOVA, per se.

If we have only two levels of our categorical explanatory variable, then retaining or rejecting the overall null hypothesis, is all that needs to be done in terms of hypothesis testing.

But if we have 3 or more levels $(k \geq 3)$, then we usually need to follow-up on rejection of the overall null hypothesis with more specific hypotheses to determine for which population group means we have evidence of a difference.

The null hypothesis (cont.)

The overall for one-way ANOVA with k groups is H_0: $\mu_1 = \ldots = \mu_k$. The alternative hypothesis is that "the population means are not all equal".

The F statistic (ratio)

we use the “F-statistic“ with $ANOVA$. The single formula for the F-statistic that is shown in most textbooks is quite complex and hard to understand. But we can build it up in small understandable steps.

Remember that a sample variance is calculated as SS/df where SS is “sum of squared deviations from the mean" and df is “degrees of freedom"

In **ANOVA** we work with variances and also “variance-like quantities" which are not really the variance of anything, but are still calculated as SS/df.

We will call all of these quantities **mean squares** or MS. i.e., $MS = SS/df$, which is a key formula that you should memorize.

Note that these are not really means, because the denominator is the df, not n.

For one-way ANOVA we will work with two different MS values called “mean square within-groups", $\mathrm{MS}_{\mathrm{within}}$, and “mean square between-groups", $\mathrm{MS}_{\mathrm{between}}$.

We know the general formula for any MS, so we really just need to find the formulas for $\mathrm{SS}_{\mathrm{within}}$ and $\mathrm{MS}_{\mathrm{between}}$, and their corresponding df.

The F statistic denominator: MS_{within}

$\mathrm{MS}_{\mathrm{within}}$ is a “pure" estimate of σ^2 that is unaffected by whether the null or alternative hypothesis is true.

For an individual group, i, $SS_i = \sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_i\right)^2$ and $df_i = n_i - 1$

We can use some statistical theory beyond the scope of this course to show that in general, $\mathrm{MS}_{\mathrm{within}}$ is a good (unbiased) estimate of σ^2 if it is defined

$$MS_{within} = SS_{within}/df_{within}$$

where $\mathrm{SS}_{\mathrm{within}} = \sum_{i=1}^{k} SS_i$, and $\mathrm{df}_{\mathrm{within}} = \sum_{i=1}^{k} \mathrm{df}_i = \sum_{i=1}^{k}(n_i - 1) = N - k.$

The F statistic ratio

It might seem that we only need $\mathrm{MS}_{\mathrm{between}}$ to distinguish the null from the alternative hypothesis, but that ignores the fact that we don't usually know the value of σ^2 . So instead we look at the ratio

$$F = \frac{\mathrm{MS}_{\mathrm{between}}}{\mathrm{MS}_{\mathrm{within}}}$$

to evaluate the null hypothesis. Because the denominator is always (under null and alternative hypotheses) an estimate of σ^2 (i.e., tends to have a value near σ^2), and the numerator is either another estimate of σ^2 (under the null hypothesis) or is inflated (under the alternative hypothesis), it is clear that the (random) values of the **F-statistic** (from experiment to experiment) tend to fall around 1.0 when

The F statistic ratio (cont.)

the null hypothesis is true and are bigger when the alternative is true. So if we can compute the sampling distribution of the F statistic under the null hypothesis, then we will have a useful statistic for distinguishing the null from the alternative hypotheses, where large values of F argue for rejection of H_0.

The F-statistic, defined by $\boldsymbol{F = \frac{MS_{between}}{MS_{within}}}$

tends to be larger if the alternative hypothesis is true than if the null hypothesis is true.

Test statistic:

$$F = \frac{s_B^2}{s_W^2} \overset{H_0}{\sim} F_{t-1,n_T-t}$$

$$\text{where} \quad s_B^2 = \frac{SSB}{t-1}, \quad s_W^2 = \frac{SSW}{n_T-t}.$$

$$\text{Here} \quad SSB = \sum_{i=1}^{t} n_i \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2 \quad \text{(Sum of squares between samples);}$$

$$SSW = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 \quad \text{(Sum of squares within samples)}$$

$$= TTS - SSB \quad ;$$

$$TSS = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 \quad \text{(Total sum of squares);}$$

y_{ij}: The jth sample observation selected from population i. For example, y_{23} denotes the third sample observation drawn from population 2.

n_i: The number of sample observations selected from population i. In our data set, n_1, the number of observations obtained from population 1, is 4. Similarly, $n_2 = n_3 = n_4 = n_5 = 4$. However, it should be noted that the sample sizes need not be the same. Thus, we might have $n_1 = 12$, $n_2 = 3$, $n_3 = 6$, $n_4 = 10$, and so forth.

y_i: The sum (total) of the sample measurements obtained from population i.

$y_{..}$: The sum (grand total) of *all* sample observations: $y_{..} = \sum y_{i.}$.

$\bar{y}_{i.}$: The average of the n_i sample observations drawn from population i, $\bar{y}_{i.} = y_i/n_i$.

$\bar{y}_{..}$: The average of all sample observations; $\bar{y}_{..} = y_{..}/n_T$.

ANOVA Table

Source	Sum of Squares	Degrees of Freedom	Mean Square	F Test
Between samples	SSB	$t-1$	$s_B^2 = SSB/(t-1)$	s_B^2/s_W^2
Within samples	SSW	$n_T - t$	$s_W^2 = SSW/(n_T - t)$	
Totals	TTS	$n_T - 1$		

THANKS FOR LISTENING