भारतीय वन्यजीव संस्थान
Wildlife Institute of India

WII e- NEWSLETTER

Winter Edition | 2025

Volume 32 | Issue IV

Is ANOVA relevant?

- Shivam Shrotriya

At a job interview, I was once asked whether I could teach One-way ANOVA to undergraduate and post-graduate students. I replied that I could teach regression analyses. The panellist insisted that the question was specifically about One-way ANOVA. Therefore, I tried to elaborate my response that I was comfortable teaching generalized linear modelling, which includes One-way ANOVA, Two-way ANOVA, and MANOVA. To which the interviewer gave a startled look, as though I were uttering what rubbish. Another ANOVA moment happened a couple of years ago, when I was scheduled for a class on basic statistics alongside another tutor. My co-tutor requested me if I could teach ANOVA, because they'll take the class on correlation and would be more comfortable teaching regression as a "continuation". I obliged to this amusing request, teaching ANOVA as a particular configuration of the regression model. But the episode made me ponder how and why introductory statistics curricula, particularly in life sciences, forestry, wildlife biology, and environmental sciences, continue to treat ANOVA as conceptually distinct from regression, following the archaic dichotomy of *tests of difference vs tests of relationship*.

As a student, several books that I came across on statistics used this dichotomy to structure their content. Some of the known examples include Zar's *Biostatistical Analyses* (1984), Dytham's *Choosing and Using Statistics*: *A Biologist's Guide* (2003), and Hawkins' *Biomeasurement: A Student's Guide to Biological Statistics* (2009). Naturally, teachers also tend to follow this inherited structure, where statistical methods are divided into two broad categories: *Tests of Difference* (e.g., t-tests and ANOVA) and *Tests of Relationship* (e.g., correlation and regression).

In this categorising framework, *tests of difference* are presented as tools to determine whether two or more groups are "different," usually by comparing their means. For example, suppose a researcher wants to know whether the application of a pesticide increases apple yield. They select 10 trees without pesticide and 10 trees with pesticide application. The yields from these two groups are then compared using a t-test to assess whether the average yield differs significantly between treatments. ANOVA comes into play if there are more than two such groups, let's say trees without pesticide, trees with pesticide A, and trees with pesticide B. ANOVA evaluates whether the means of three or more groups differ by testing the null hypothesis that all group means are equal.

In contrast, *tests of relationship* are introduced as methods to evaluate whether two variables are associated. For instance, instead of comparing two or three discrete groups, a researcher may measure pesticide dosage continuously across trees and examine whether yield increases with increasing dosage. In that case, correlation or regression would be used to test whether a relationship exists between pesticide concentration and crop yield. This pedagogical separation creates an impression that "difference" and "relationship" are fundamentally distinct statistical problems, creating a false separation.

In ANOVA, the procedure is to calculate F-statistics, which basically compares the variance within each dataset to the variance between the datasets:

$$F = \frac{\text{Variance Between Groups}}{\text{Variance Within Groups}} \quad \text{......} \quad \text{Eq.1}$$

If between-group variability is significantly larger than within-group variability, we conclude that at least one group mean differs significantly from the others. The *variance* itself is calculated using the *mean squares* method. For within-group variation, we compute the squared difference between each observation and its group mean. For between-group variation, we compute the squared difference between each group mean and the overall mean. Squaring differences primarily prevents positive and negative differences from cancelling each other out.

Mathematically, this is similar to fitting a regression line. Mean squares are calculated for a regression line, finding the difference between the observation and the regression line, also called residuals (Figure 1). The best-fit line is the one for which the squared sum of these differences is the least, or the least squares.

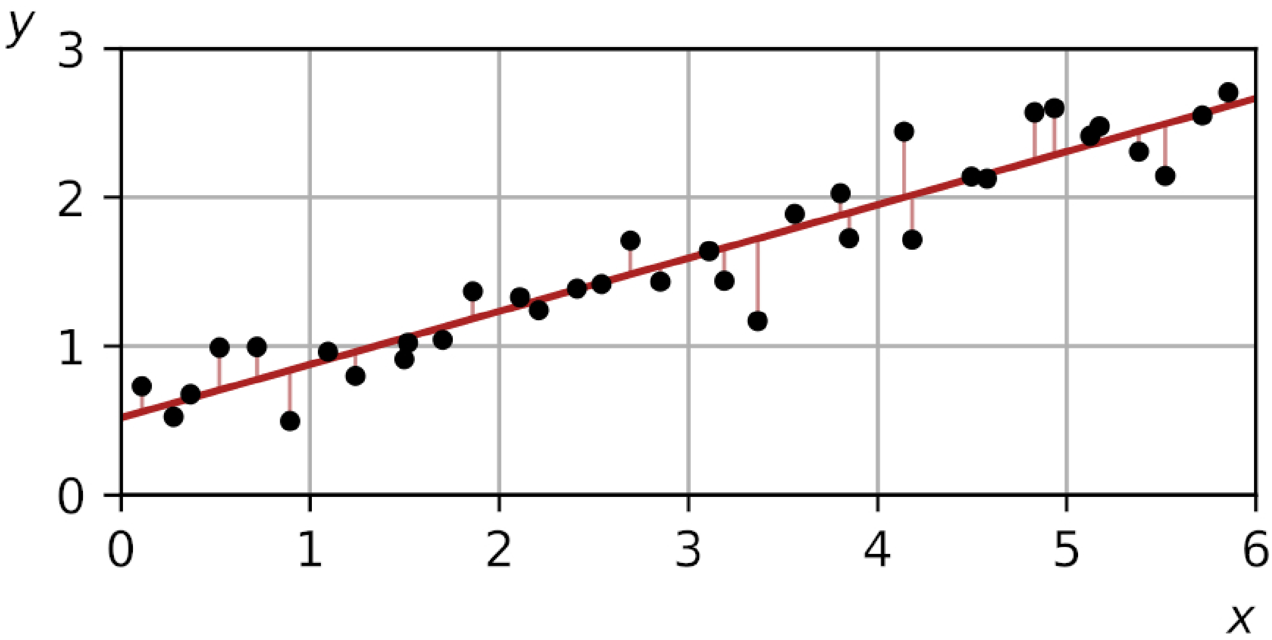

Figure 1. A best-fit regression line also has the least sum of squared residuals
(Source: Wikipedia, By Justinkunimune - Own work, CC0)

We are familiar with the mathematical expression for a simple linear regression, such as figure 1:

$$Y=\beta_0+\beta_1 X+\varepsilon \quad \text{Eq. 2}$$

where, Y is the response variable, X is the predictor variable, β_0 is the intercept (or where the line will begin on the y-axis), β_1 is the slope of the line, is the unaccounted effect that is causing the observations to deviate from the line.

Let's suppose X is not continuous but categorical. In the pesticide example with three treatment groups, we could present it as a regression where the response variable is yield across all samples and the predictors are the treatment groups as dummy variables, or each category turned into presence/absence (1,0) data (Figure 2).

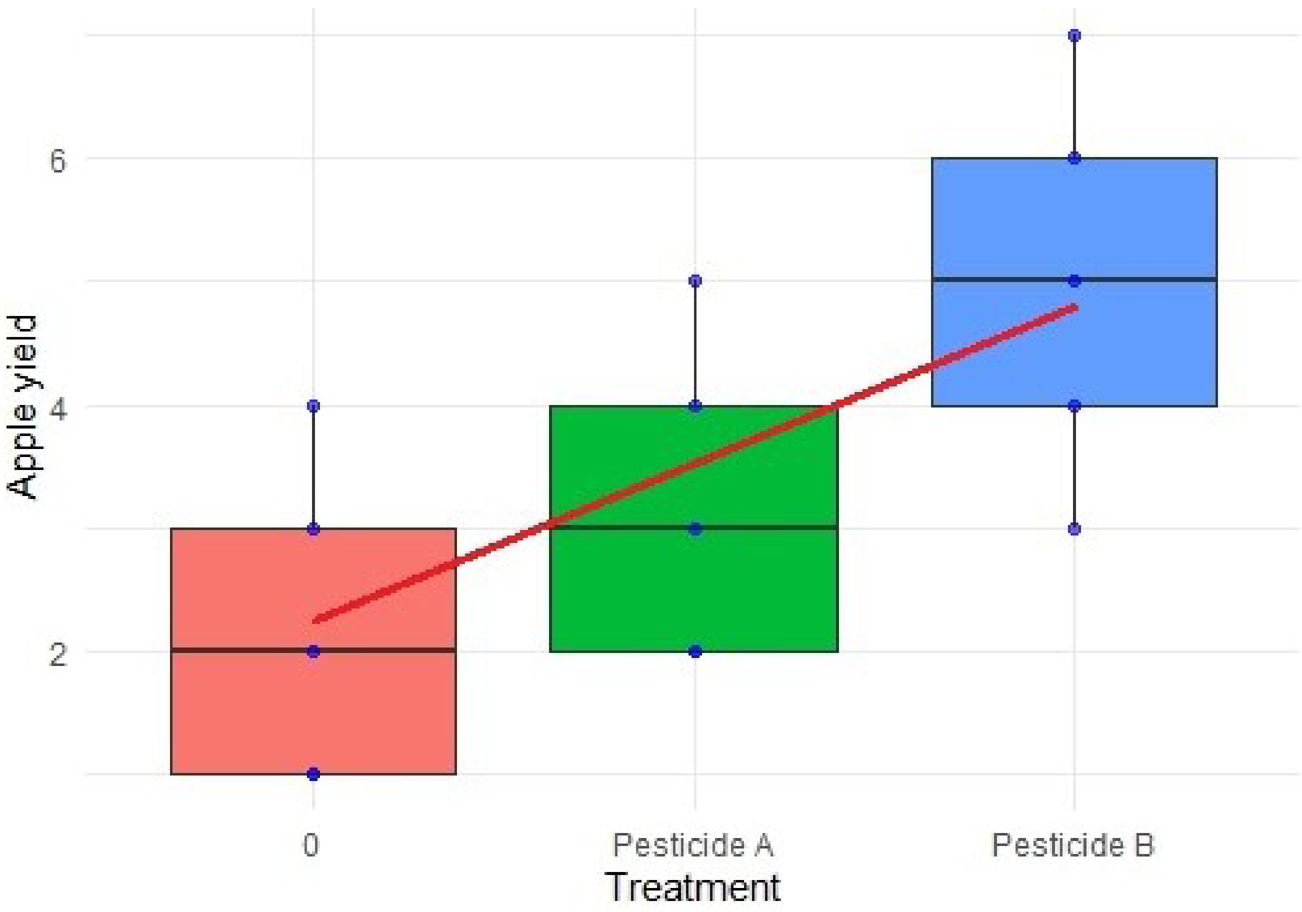

Figure 2. ANOVA can also be presented as a regression setup when the predictor variable is categorical.

The mathematical expression of this linear regression would be:

$$Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\varepsilon \quad \text{Eq. 3}$$

where, X_1 and X_2, both are pesticide treatments. This is simply one-way ANOVA written as a regression. The total sum of squares (SS) in ANOVA is partitioned exactly as in regression:

$$\textit{Total SS}=\textit{Model SS}+\textit{Residual SS} \quad \text{Eq. 4}$$

The F-statistic computed from the regression model is numerically identical to the ANOVA F-statistic. The p-value is the same. The phrasing of the statistical null hypothesis in ANOVA is that *"all group means are equal"*, which could be rephrased for regression as *"all group coefficients (except intercept) are zero."* In fact, in the words of Andy Field, "all statistical tests boil down to variants on regression", and "ANOVA is just a special case of regression" (Field *et al*. 2012). Cohen (1968) also demonstrated decades ago that most classical statistical procedures, including t-tests and ANOVA, are special cases of the general multiple regression model. F-statistic, as the variance-ratio, is simply a computational shortcut historically developed before matrix algebra and computing became accessible.

The unified framework here at work is the General Linear Model (GLM):

$$Y=X\beta+\varepsilon \quad \text{Eq. 5}$$

where, Y is the response vector, X is the design matrix, containing numeric and/or dummy (1,0) predictor variables, β is the coefficients, and ε is the residuals.

Now the question arises, if ANOVA is essentially a regression, how did the two get separated in textbooks and in practice? The roots of this separation can be traced to the long-standing disciplinary split between experimental and correlational sciences (Cronbach, 1957). Experimental research adopted ANOVA as a go-to analytical tool, while observational and correlation research leaned towards regression as the primary method. Methodological camps evolved in parallel rather than in integration, leading to a statistical split. Textbooks and curricula followed suit, and students began learning ANOVA and regression in entirely different contexts (Field *et al*. 2012).

Field *et al*. (2012) note that explaining ANOVA via variance ratios works for simple designs but becomes cumbersome for complex situations, such as ANCOVA, unequal sample sizes, and interactions between treatments. The regression framework extends naturally to these cases. One-way ANOVA is regression with one categorical predictor, Two-way ANOVA is regression with two categorical predictors and interaction, and MANOVA is a special case of the multivariate linear model:

$$Y=XB+E \quad \text{Eq. 6}$$

where, Y is a matrix of multiple dependent variables, X is a design matrix (usually of categorical group indicators), E is a matrix of residuals, and inference focuses on testing hypotheses about group mean vectors.

ANCOVA takes the shape of a regression with categorical treatments with continuous covariates:

$$Y= \beta_0+\beta_1\, Treatment+ \beta_2 Covariate+ \varepsilon \quad \text{Eq. 7}$$

Analysis of Molecular Variance (AMOVA) in population genetics partitions genetic variance across hierarchical levels. Conceptually, it can be presented as a hierarchical linear model:

$$Y= \mu +Population+Individual(Population)+ \varepsilon \quad \text{Eq. 8}$$

where, μ is the overall mean genetic distance, and the observed genetic distances (Y) are modelled as the deviation of each population mean from the overall mean and the deviation of individuals from their population means.

Further, carrying out ANOVA still remains work half done, as it can only tell that there is at least one group that differs from the others. But identifying the specific group-pairs with significant differences requires further post-hoc tests or running planned contrasts (see Field *et al*. 2012). In regression, each β represents the effect of a group relative to a reference category, confidence intervals indicate whether that effect differs from zero, and contrasts can be directly specified within the model. Regression treats group differences as estimated parameters rather than secondary discoveries.

The separation between *tests of difference* and *tests of relationships* persists not because of conceptual necessity, but because of historical inertia. Learning ANOVA within the regression framework provides conceptual continuity. Students who already understand regression slopes and residuals can extend that knowledge to categorical predictors. It also shifts focus from p-values to effect sizes. Regression emphasizes coefficients, magnitudes, and confidence intervals, which are biologically more meaningful. The understanding now expands from just whether something differs to by how much. Moreover, the regression framework scales logically to complexity. Interactions, covariates, unequal designs, and hierarchical structures are handled by expanding the design matrix, without introducing new "tests" for different setups. In fact, statistical programs like R handle ANOVA as post-processing of regression, rearranging the regression outputs in a way ANOVA users are familiar with. Most importantly, regression-based learning encourages model-based thinking. Instead of asking, "Which test should I use?" the researcher can focus on, "What biological/ecological process am I modelling? What variables influence my response? How do I structure the model?"

ANOVA, as an analytical framework, has played an important role in the history of ecology. For example, classical studies by Robert H. MacArthur on niche partitioning (MacArthur, 1958), Joseph H. Connell on competition (Connell, 1961), and Robert T. Paine on keystone species (Paine, 1966) used ANOVA-type study design and analysis to test experimental hypotheses, which gave clear inferences leading to significant theoretical advances in ecology. However, for these relatively simple and controlled designs, ANOVA was sufficient and powerful, but modern ecological data analysis increasingly relies on unified modelling frameworks (Quinn & Keough, 2002).

Since these early works, ecological research has grown more complex with multivariate systems becoming the norm. Our studies more often include multi-species communities, hierarchical processes, spatial autocorrelation, and non-normal data structures. Statistical ecology more frequently uses generalized linear models, mixed models, Bayesian approaches, and hierarchical modelling. The continued separation of *tests of difference* and *tests of relationship* in the introductory statistics classes, though pedagogically convenient, obscures their shared mathematical foundation. Therefore, we need a coherent understanding of how these statistical procedures function, especially in the context of modern ecology.

The problem is not merely technical and terminological, or 'what is in the name', if it is essentially the same. How statistics is taught shapes how students frame research questions. Viewing statistical analysis as choosing from a list through a selection system of distinct and disconnected tests leads to procedural matching. In understanding the unified modelling framework, focus shifts to hypothesis structure, predictor variables, variance partitioning, and biological mechanism. We need to teach model construction instead of test selection and memorizing procedures. In doing so, students will not just learn statistics, they will grow a statistically aware mindset.

And perhaps the next time someone is asked whether they know ANOVA, it can be confidently replied with "Yes, as a regression model."

Acknowledgement:
The author thanks Dr. Akanksha Saxena for her comments on the draft of this article.

Key references:

- Cohen, J., (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70(6):426-443.
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, 12:671–684.
- Field, A. *et al.* (2012). *Discovering statistics using R*. SAGE publications.
- Quinn, G. P., & Keough, M. J. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge University Press.

AUTHOR:

Dr. Shivam Shrotriya *is Assistant Managing Editor at the Journal of Wildlife Science and holds a PhD on the ecology of the Himalayan wolf. He is passionate about all thing's ecology. He reads for work and for fun. He sometimes teaches population ecology and analytical methods for different classes at the Wildlife Institute of India. In this column, he revisits classical papers and foundational theories in wildlife ecology and conservation science with fresh eyes -exploring how new insights, data, and perspectives have reshaped these concepts.*

Bluesky: @shivamshrotriya.bsky.social
Insta: @shivamshrotriya
Twitter: @shivam_wolf