Assessing the Assumptions of the Analysis of Variance

Norberto E. Milla, Jr.

2025-09-08

ANOVA Assumptions

Recall: (*Effects model*)

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, i = 1, 2, \ldots, t; j = 1, 2, \ldots, r_i$$

- $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$
- Under the fixed effects model,τ_i are fixed quantities with $\sum_{i=1}^{t} \tau_i = 0$
- Under the random effects model, τ_i are random quantities with $\tau_i \sim N(0, \sigma^2_\tau)$

ANOVA Assumptions

- The experimental errors are normally distributed (**Normality**)
- The experimental errors have equal variance (**Homogeneity of variance**)
- The experimental errors are independent (**Independence**)
- Treatment effects and environmental (random) effects are additive (**Additivity**)

Assessing ANOVA Assumptions

- the quality of our inference depends on how well the errors ϵ_{ij} conform to the assumptions
- decisions about how well the errors meet our assumptions are based not on the errors ϵ_{ij} themselves, but instead on observed residuals e_{ij}
- in any real-world data set, we are almost sure to have one or more of the three assumptions be false, e.g. real-world data are never exactly normally distributed
- reasonable inferences can still be reached when the departures from our assumptions are not too large (robustness)
  - the F test and interval estimates of effects are reliable even if the assumptions are not met however, this robustness is very difficult to quantify and is also very dependent on balanced sample sizes
  - the F test can become very unreliable when unequal sample sizes are combined with non-normal data with heterogeneous variances

Assessing ANOVA Assumptions

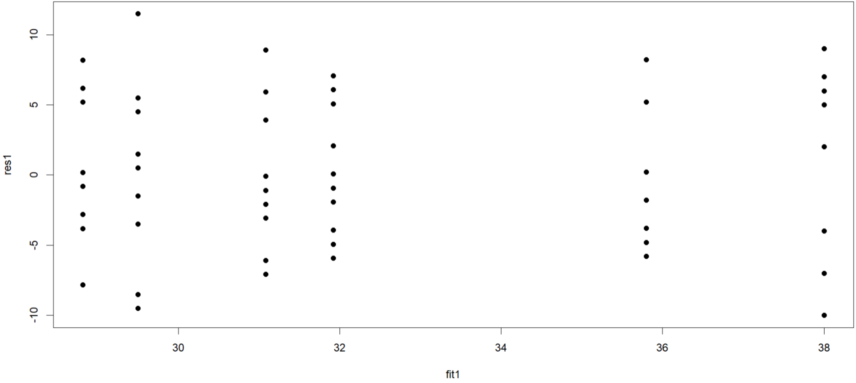
- Experimental error is approximately normally distributed when the sample sizes are adequately large
  - When sample sizes are small (which is very common in most experiments), normality of errors is in doubt
- A large deviation from normality leads to hypothesis test conclusions that are too liberal and a decrease in power and efficiency
- Normality is oftentimes violated when:
  - Error variances are heterogeneous
  - Data are extremely skewed, especially when there are outliers

Tests for normality of errors

- Graphical approach: Normal probability plots, histogram, boxplot
- Formal tests:
  - Wilk-Shapiro W test ($10 \leq n \leq 50$)
  - Kolmogorov-Smirnov test ($n > 50$)
  - Anderson-Darling Test
  - Cramer-von Mises Test
- H_0: The errors are normally distributed.
- H_1: The errors are not normally distributed.

Detecting variance heteroscedasticity

- Graphical approach
  - Draw a plot of the residuals (Y-axis) against the fitted values (X-axis).
  - If the variance is constant, the vertical spread in the stripes will be about the same



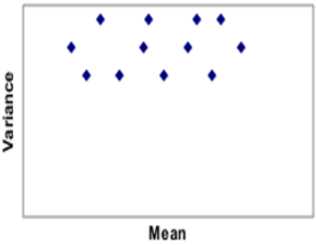
Types of variance heteroscedasticity

- Variance is functionally related to the mean
  - Associated with variables that are not normal such as counts—number of leaves, number of lesions, number of eggs, number of infected cells (Poisson distributed)
  - Survival data (binomial)—the variance is usually of the form

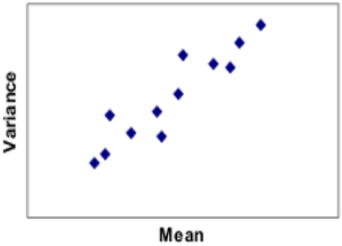
$$S^2 = \overline{X}\left(1 - \overline{X}\right)$$

- No functional relationship between variance and mean
  - Due to the nature of treatments—some treatments have errors that are substantially higher than others
  - Presence of outliers

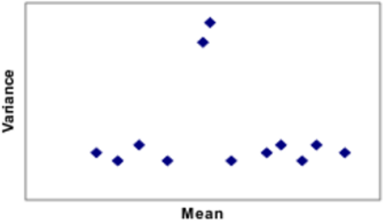
Types of variance heteroscedasticity



Homogeneous variance



Variance proportional to mean



No functional relationship between mean and variance

Causes of heterogeneity of error variances

- Experimental error variance is not constant across groups or populations (heteroscedasticity)
- Some treatments are erratic in their effects
- Presence of outliers
- Nonnormal or skewed distribution of data

Tests for homogeneity of error variance

- $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2$
- $H_1 : \sigma_i^2 \neq \sigma_i^2$, for any $i \neq j$
- Test statistics:
  - Hartley's F-max test
  - Levene's test
  - Bartlett's test
  - Brown and Forsythe's test
  - Fligner-Killeen test
  - O'Brien's test

Some remarks

- Both Hartley's F-max test and Bartlett's test are quite sensitive to departures from normality as well as to departures from the equal variances assumption
- Levene's test, Brown and Forsythe's test and O'Brien's test are robust to departures from normality
- Simulation studies suggests the following:
    - If the distributions have heavy tails, use the Brown and Forsythe's test
    - If the distribution is somewhat skewed, use O'Brien's test
    - If the data are nearly normally distributed, then any of the tests can be used

Independence of errors

- Possible reasons why independence of observations may be violated
  - responses on neighboring plots tend to be correlated
  - adjacent plants have some form of symbiotic relationship or competition for light, water and other environmental factors
  - neighboring households are closely related
  - experiments where experimental units or replications are arranged systematically
- Positively correlated data inflates standard error
- The remedy is proper randomization of treatments to the experimental units

Independence of errors

- Graphical Approach
  - Plot the fitted values against the residuals
  - If no definite pattern, then the errors can be taken to be independent
  - If with upward/downward pattern, then the errors are correlated
- Formal Tests
  - Correlation between adjacent residuals
  - Runs test
  - Durbin Watson test (time series data)

Variance-stabilizing transformations

- Logarithmic transformation
  - Appropriate for data where its standard deviation is proportional to its mean as in data with greatly skewed distribution
  - Appropriate for data with multiplicative effects
  - Form: $Y_{ij}^* = log(Y_{ij})$
  - When data contains 0's and very small values, use $Y_{ij}^* = log(Y_{ij} + 1)$

Variance-stabilizing transformations

- Square root transformation
  - Appropriate for data where its variance is proportional to its mean as in data with small whole numbers like counts of rare events (Poisson data) and frequency count data
  - Form: $Y_{ij}^* = \sqrt{Y_{ij}}$
  - When data contains small values (say, less than 10) or when 0's are present use $Y_{ij}^* = \sqrt{Y_{ij} + 0.5}$
- Arcsine transformation
  - Appropriate for proportion or percentage data (binary data): $Y_{ij}^* = \arcsin(\sqrt{Y_{ij}})$ or $Y_{ij}^* = \arcsin(\sqrt{Y_{ij} + 0.5})$

Fixing Problems

- Nonnormality
  - transform data to another scale
  - skewness to the right is lessened by a square root, logarithm, or other transformation to a power less than one
  - skewness to the left is lessened by a square, cube, or other transformation to a power greater than one
  - symmetric long tails do not easily yield to a transformation
  - Robust and rank-based methods can also be used in cases of non-normality
  - outliers can affect analysis
  - perform the analysis both with the full data set and with outliers excluded

Fixing Problems

- Nonnormality
  - if conclusions change when the outliers are excluded, then you must be fairly careful in interpreting the results
  - outliers need not be bad data points; in fact, they may be the most interesting and informative data points in the whole data set
  - they just don't fit the model which probably means that the model is wrong
- Heterogeneous variances
  - Apply variance-stabilizing transformation
  - Apply weighted ANOVA (e.g. Welch test)
- Dependent errors
  - If serial correlation (autocorrelation) is present, include in the model a term that will account time trend