The effects of host environment on the growth of a butterfly species

Márton Horváth

2023-11-11

Introduction

Analysis of Variance (ANOVA) is a powerful statistical technique used to examine the differences between multiple groups or treatments and determine whether those differences are statistically significant. ANOVA is particularly valuable when you want to compare means across more than two groups, making it a versatile tool in various fields, including psychology, biology, business, and more. In this exercise, I apply ANOVA to a real-world data set to explore the sources of variation on the growth of a butterfly species. In this case the possible contributors are the plant hosts that were available at different stages of the life cycle of the butterfly.

```
##   LarvalID LarvalHost Sex MaternalHost MotherID DevelopmentTime AdultWeight GrowthRate
## 1       14   Barbarea   M     Barbarea        9              21          64      0.086
## 2       22   Barbarea   M     Barbarea        9              22          61      0.081
## 3       23   Barbarea   F     Barbarea        9              22          53      0.078
## 4       24   Barbarea   M     Barbarea        9              22          68      0.083
## 5       25   Barbarea   F     Barbarea        9              22          57      0.080
## 6       30   Barbarea   F     Barbarea        9              23          89      0.085
```

Methods

Normality of the data

ANOVA assumes that the data is normally distributed, thus I used the Shapiro-Wilk test to determine whether the three possible response variables: the ‘development time’, the ‘growth rate’ and the ‘adult weight’ follow a normal distribution or not. This statistical test assesses the null hypothesis that the data is normally distributed. If the p-value from the test is less than the chosen significance level (0.05), the null hypothesis can be rejected, indicating that the data significantly deviates from a normal distribution. The test was run on the residues from the mixed-effect linear model regression `lm(DevelopmentTime ~ LarvalHost * MaternalHost, data = data, na.action = na.exclude)`. I have go the standardized residues from the model using `rstandard()`, then used `saphiro.test()` function to perform the test. It’s important to note that the Shapiro-Wilk test is most suitable for relatively small sample sizes (typically less than 50 observations). For larger data sets, the test can become overly sensitive and may detect minor deviations from normality that are not practically significant. Hence, two visual methods were also included assessing the normality alongside the formal statistical test. Histograms were created for each response variable and Q-Q (quantile-quantile) plots were drawn on the residues.

Homogeneity of variance

The ANOVA also assumes homogeneity of variance, which means that the variance among the groups should be approximately equal. Levene’s test was used to check if the variances whether the response variables obtained for the four groups are equal or not. when data comes from a non-normal distribution. Levene’s test was used to check the null hypothesis that the population variances are equal, also known as homoscedastic.

Analysis of variances

Analysis of Variance (ANOVA) was used to analyze and compare the means of the distinct groups, to test whether there are statistically significant differences between them. In this case two-way ANOVA was employed to analyze the two independent variables - ‘larval host plants’ and ‘maternal host plants’ - and their interaction. ANOVA generates an F-statistic, which measures the ratio of between-group variance (MSB) to within-group variance (MSW) - $F = \frac{MSB}{MSW}$, where:

$$MSB = \frac{\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2}{K-1}$$

where:

- k is the number of groups,
- n_j is the sample size of the j-th group,
- $\bar{y}_j$ is the mean of the j-th group,
- $\bar{y}$ is the overall mean.
- and K is the number of groups

$$MSW = \frac{\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2}{N-1}$$

where:

- n is the total number of observations,
- y_{ij} is an individual data point in the j-th group,
- $\bar{y}_j$ is the mean of the j-th group.
- and N is the total number of data points.

Understanding the results of the ANOVA analysis

The summary of the ANOVA output doesn’t provide information about which specific groups differ from each other. Tukey’s Honestly Significant Difference (HSD) post-hoc test was used to determine the pairwise differences and to reveal which identify which group comparisons are responsible for these differences. The interpretation of the results of the Tukey’s HSD test two select the response variable with the biggest effect size.

Results

1.) Only the adult weight of the butterflies is suitable for ANOVA

Both the Shapiro-Wilk test and Levene’s test showed that the two response variables: ‘development time’ and ‘growth rate’ do not satisfy the criteria for ANOVA analysis (p-values: p=1.9e-06 and p=3e-02, and p=4e-05 and p=1.9e-04, respectively). The violation of the normality of the data can be seen on both the histograms (fig 1.) and the Q-Q plots (fig 2.), the resulting lines showed discrete distribution with positive skew and negative skew alongside with truncation, respectively. For these non-normally distributed data a non-parametric test, such as the Kruskal-Wallis or Friedman tests, could have been used. Additionally, violation to the homogeneity of variances, could have been mitigated using a Welch ANOVA, however, both of these methods fall outside of the scope of the current analysis. Only the response variable ‘adult weight’ passed both the Shapiro-Wilk and Levene’s tests (p=3.3e-01 and p=2.1e-01, respsectively), hence it was used during the further analysis steps.

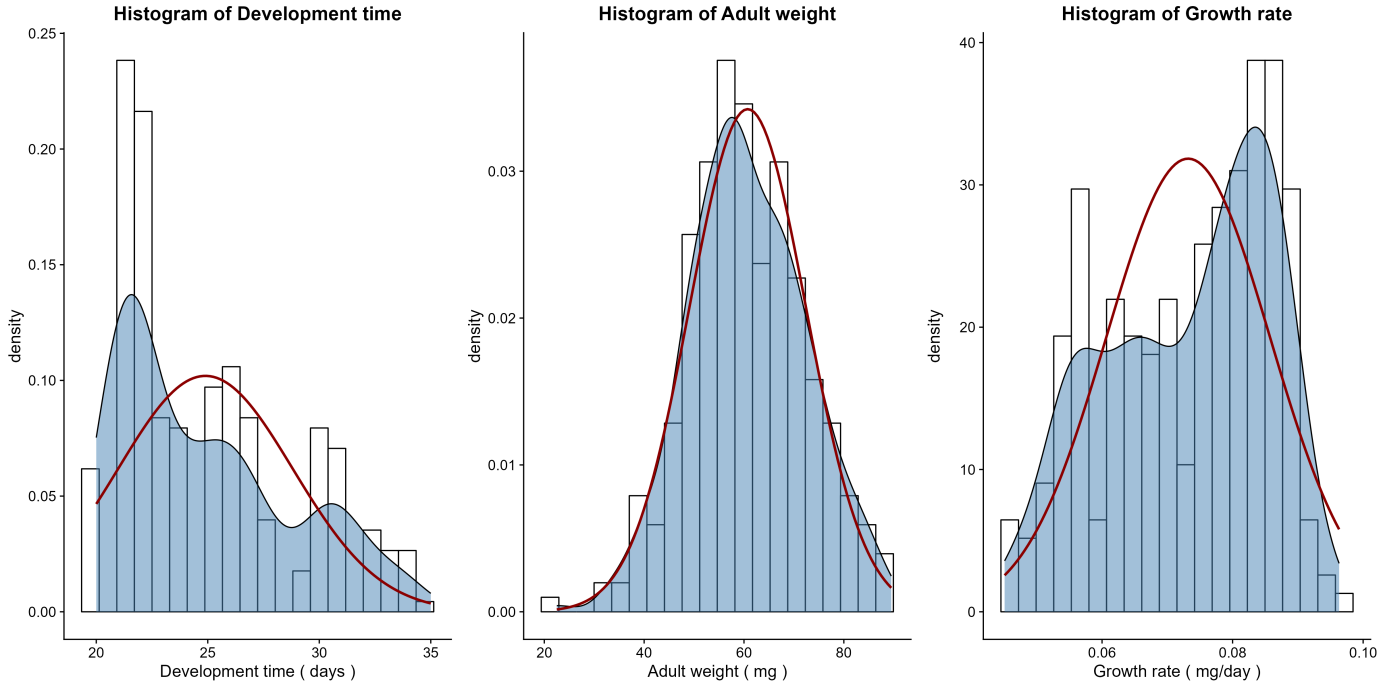
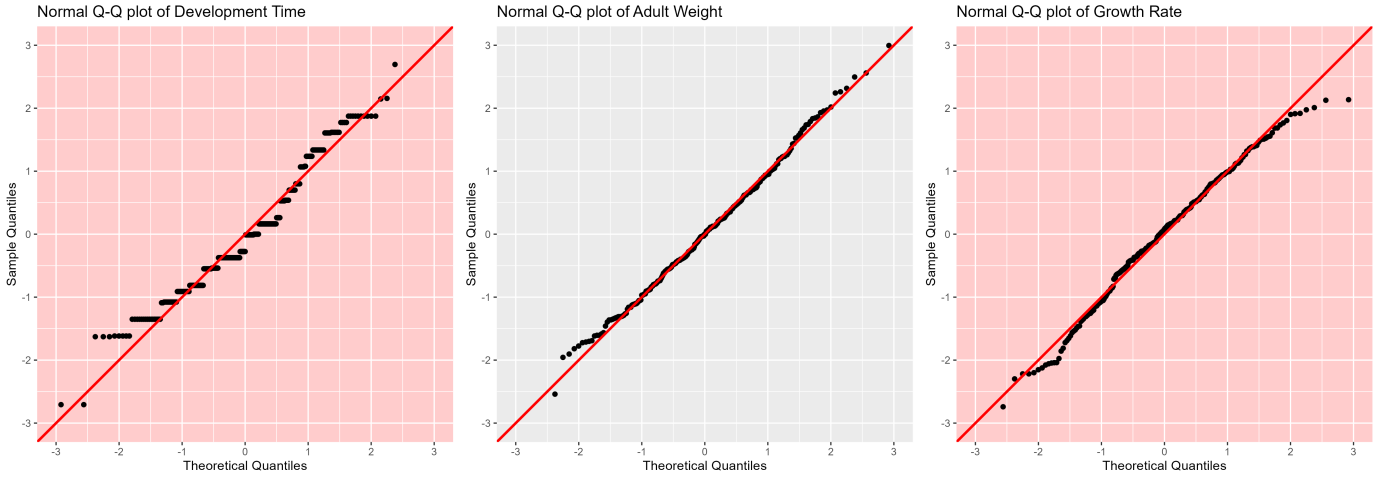


Fig 1.) Histogram of the measurement points in each of the three response variables, with the theoretical normal distribution (red) and the actual distribution of the data (blue).



Mean	SE	Median	Skew	Kurtosis
0	0.006	-0.513	0.555	0.329

Mean	SE	Median	Skew	Kurtosis
0	0.033	-0.024	0.012	0.406

Mean	SE	Median	Skew	Kurtosis
0	0	0	-0.336	-0.114

Fig 2.) Q-Q plot of the residues, the background is highlighted red if the variable failed the Saphiro-Wilk test. Descriptors, such as the mean, statistical error, median, skewness and kurtosis values of the data are included under each plot.

2.) The larval host plant alone determines the adult weight of the butterflies

The ANOVA results suggest that there is a significant difference between the groups and the post-hoc test showed that the variance mainly depends on the larval host plant.

```
##                                          diff       lwr       upr    p adj
##
## Berteroa-Barbarea                    -1.0e+00 -3.2e+00  1.2e+00  3.7e-01
## Berteroa-Barbarea                    -1.4e+01 -1.6e+01 -1.2e+01  4.9e-13
## Berteroa:Barbarea-Barbarea:Barbarea   1.4e+00 -2.3e+00  5.1e+00  7.7e-01
## Barbarea:Berteroa-Barbarea:Barbarea  -1.2e+01 -1.6e+01 -7.8e+00  6.4e-12
## Berteroa:Berteroa-Barbarea:Barbarea  -1.5e+01 -1.9e+01 -1.1e+01  5.3e-13
## Barbarea:Berteroa-Berteroa:Barbarea  -1.3e+01 -1.8e+01 -8.9e+00  8.5e-13
## Berteroa:Berteroa-Berteroa:Barbarea  -1.6e+01 -2.1e+01 -1.2e+01  5.3e-13
## Berteroa:Berteroa-Barbarea:Berteroa  -3.1e+00 -7.9e+00  1.6e+00  3.2e-01
```

The ANOVA (`formula: AdultWeight ~ MaternalHost * LarvalHost`) suggests that the effect of the maternal host is statistically not significant and very small (F(1, 283) = 0.82, p = 0.366, Eta2 (partial) = 2.89e-03), so is the interaction between maternal host and larval host (F(1, 283) = 3.75, p = 0.054, Eta2 (partial) = 0.01), however, the effect of larval host is statistically significant and large (F(1, 283) = 144.89, p < .001, Eta2 (partial) = 0.34). The ANOVA results showed that the combined effect of the maternal- and larval host plants is responsible for the 34.5597022343403 % of the total variation that was observable in the weight of the adult butterflies. However, type of the plant that the larvae fed on alone was responsible for 97.44 % of this effect. It can be concluded that the most important factor on the growth of the butterfly is the host plant during the larval phase of their life cycle.

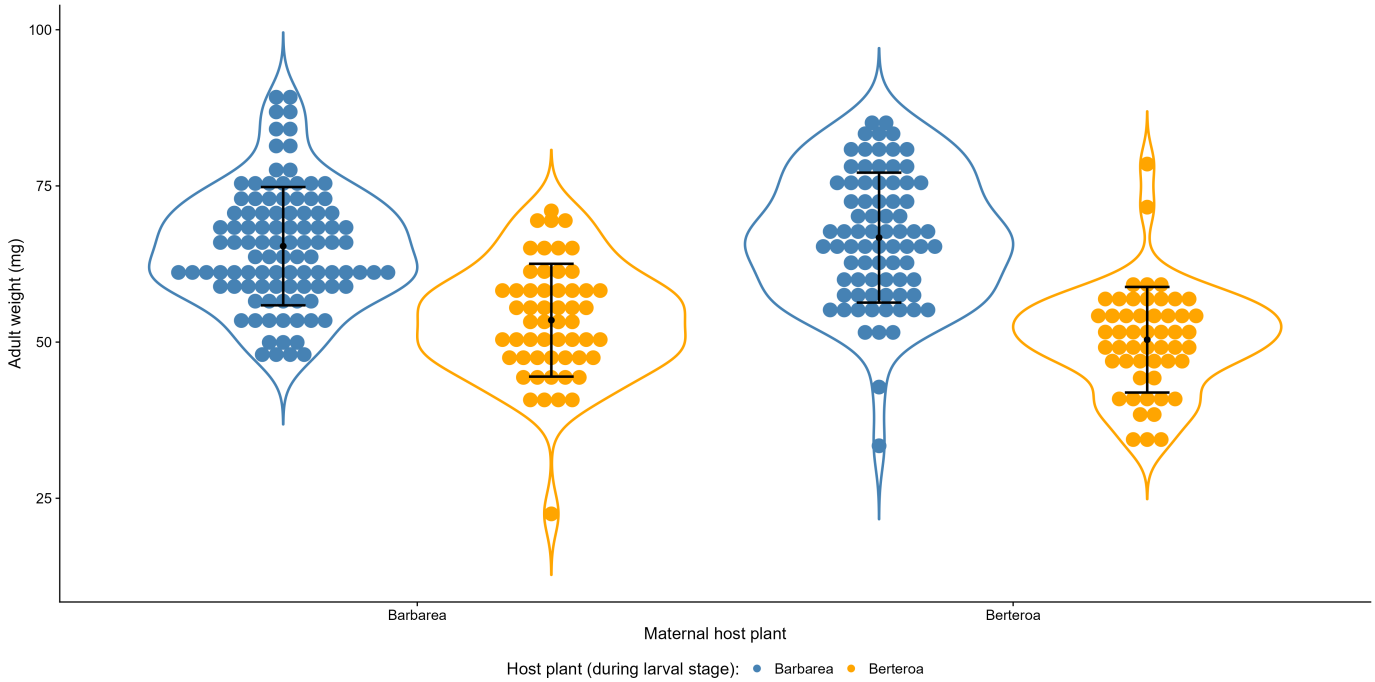


Fig 3.) Violint plot showing the difference in the mean weight of the adult butterflies depending on the available food during different stages of their lify cycle.