DS805: Multivariate Statistical Analysis
University of Southern Denmark - IMADA
June 2025

Leila Sabanovic, lesab19
Hawa Mahamed Adalla, haad24
Mathias Vestergaard, maves24
Mette Skovsen, mesko24

Anaconda Data

Introduction

- This project analyzes the statistical difference in length and weight between female and male anacondas.
- The aim is to determine whether there is an observed difference in size between female and male anacondas, and whether this difference is statistically significant.
- This analysis was performed on a dataset (T6-19)[1] containing 56 observations (28 females and 28 males) with two measurable variables (length and weight).
- As each group in our dataset contains fewer than 30 instances, we are working with a small-sample dataset. Therefore, we select methods appropriate for this.

The null hypothesis and the alternative hypothesis state that:

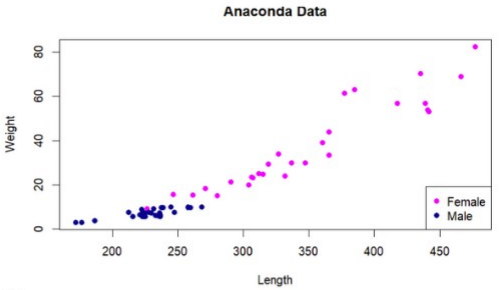
$H_0 : \mu_{\text{Male}} = \mu_{\text{Female}}$ $\quad$ $H_1 : \mu_{\text{Male}} \neq \mu_{\text{Female}}$

"Males and females do not differ in the two physical attributes, weight and length"

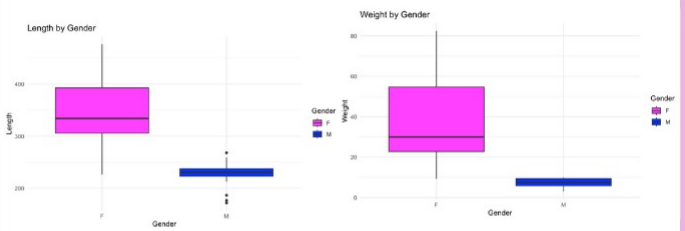
"Males and females do differ in the two physical attributes, weight and length"

Descriptive Statistics

Scatter Plot



Box Plots



- The **scatter plot** indicates a strong positive correlation between the two variables.
- It shows a clear separation between females and males, with female anacondas being longer and heavier on average, and also shows greater variation in size.

- The **box plots** show that female anacondas are both longer and heavier than males.
- The mean length both confirming consistent size differences between the sexes.

Variable	Female	Male
Length (cm)	227 cm - 477 cm	172 cm - 268 cm
Weight (kg)	9.25 kg - 82.5 kg	3.0 kg - 10.07 kg
Mean Length (cm) – SD	384.3 cm ± 68.62	228.8 cm ± 22.44
Mean Weight (kg) – SD	37.26 kg ± 20.10	7.29 kg ± 2.10

Methods

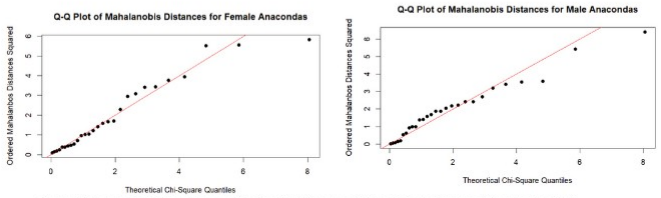
In order to test our hypothesis, we first assess the assumptions of our statisical analysis as well as performing our specific tests. We therefore:

- Assess the assumption of multivariate normality using Q-Q plots
- Assess the assumption of equality of covariance matrices using Box' M-test.
- Apply Hotelling's T^2-test to evaluate potential differences between means.
- Assessing Hotelling's T^2- and Bonferroni-adjusted confidence intervals and discriminant coefficients
- **Methodological opt-outs:**
  - Since the dataset consists of only two groups, we do not perform a MANOVA. This would be redundant, as Hotelling's T^2-test is specifically designed to compare two multivariate means. We also choose against performing a Principal Component Analysis (PCA), as having only two groups limits its practical utility in this case.

Analysis

Q-Q Plots

Our further analysis rests on the assumption of multivariate normality, so we computed Q-Q plots for a visual inspection of this:



- Generally, most of the data points for both sexes follow the reference line indicating normality .
- There are few deviations, primarily in the upper quantiles with some right tail deviations indicating some non-normality for the upper extremes.
- We determine these deviations to lie within an acceptable limit and continue with our analysis.

Box's M-Test

We evaluate equality of covariance matrices using Box' M-test, obtaining the result:

Chi-Sq (approx.) = 100.32, df = 3, p<2.2*10-16

- The very small *p*-value provides strong evidence for a rejection of the assumption of equal covariance matrices and instead indicates a significant difference between the groups
- Comparing the generalized variance, we also see that it is significantly larger for females (221533.19) compared to males (698.53), indicating a much greater variation in length and weight compared to males:

$$\Sigma_F = \begin{bmatrix} 4709.30 & 1296.76 \\ 1296.76 & 404.04 \end{bmatrix} \quad det = 221533.19$$

$$\Sigma_M = \begin{bmatrix} 503.48 & 39.13 \\ 39.13 & 4.43 \end{bmatrix} \quad det = 698.53$$

- This further supports the result of Box's M-test, which showed that the covariance matrices are significantly different between the groups.

Hotelling's T^2-Test

We also computed Hotelling's T^2-test statistic in order to compare the mean vectors:

T^2 = 76.92, p < 6.43e-11

- The large Hotelling's T^2-test statistic indicates a statistically significant difference between the two groups.
- The very small *p*-value of far less than 0.05 further reinforces that there is a statistically significant difference.

Rejection of Null Hypothesis

With the evaluation of Hotelling's T^2-test statistic and Box' M-test, we reject our null hypothesis of equal mean vectors between female and male anacondas and accept the alternative hypothesis; there is a significant difference in the average length and weight between the two groups.

Confidence Intervals and Discriminant Coefficient Vector

After having rejected the null hypotheses, we proceed by evaluating the difference in group means using 95% Hotelling's T^2- and Bonferroni-adjusted confidence intervals:

95% Hotelling's T^2 confidence intervals

Variable	Lower	Upper
Length	84.83413	154.20873
Weight	20.26326	39.68317

95% Bonferroni confidence intervals

Variable	Lower	Upper
Length	88.06248	150.98038
Weight	21.16696	38.77946

- Both confidence intervals consist solely of positive values, showing that both the length and weight are significantly greater in female anacondas.

We also assess which variable primarily enforces the observed difference by computing the coefficient vector $\hat{\boldsymbol{a}}^*$ in order to assess the relative contribution of the variables as discriminators:

Variable	Coefficient	Normalized
Length	4.0101061	0.9947753
Weight	-0.4115375	-0.1020889

- After standardizing the variables, it is revealed that the length variable accounts for the majority of the discriminating power.

Thus, while both length and weight differ in female and male anacondas, length is the component contributing the most to the difference.

Discussion

Conclusion

This analysis revealed a statistically significant multivariate difference in body size between female and male anacondas. Descriptive statistics showed that female anacondas are, on average, both longer and heavier than males. These differences were strongly supported by Hotelling's T^2 test (T^2 = 76.92, p < 6.43e-11), which demonstrates that there is a statistically significant difference between the mean vectors for the two groups.

Summing up on our results, we conclude that our findings offer clear statistical evidence for sexual size dimorphism in anacondas. This difference may be linked to biological and ecological factors such as reproductive capacity, mate selection, or energy storage needs during the gestation period.

Box's M-Test

Although Box' M-test indicated a violation of the assumption of homogeneity of covariance matrices, this is unlikely to compromise the validity of the Hotelling's T^2 result, as the group sizes are equal (n = 28). According to Rencher (Section 3.8)[2], equal group sizes mitigate the effects of unequal covariances and help maintain the robustness of the test, reducing the risk of Type I errors.

Assessment of Multivariate Normality

Regarding the assessment of the assumption of multivariate normality, we assessed this visually using Q-Q Plots. This could however also be expanded to include Mardia's Test or Henze-Zirkler's Test as well as a dedicated outliers-examination, e.g. using Mahalanobis' Distance.

Future Directions

Finally, while the statistical evidence is strong, the dataset has limitations. Only two physical variables (length and weight) were included, and the total sample size was modest (n = 56). Future work would benefit from larger and more diverse samples, including additional variables such as age, girth, offspring count, or habitat conditions. This would enhance the generalizability of the findings and open up possibilities for classification or predictive modeling.

References

1: Johnson, Richard A., and Dean W. Wichern. Applied Multivariate Statistical Analysis. 6th Pearson New International ed. Harlow: Pearson Education Limited, 2013.

2: Rencher, Alvin C. Multivariate Statistical Inference and Applications. 1st ed. New York: Wiley-Interscience, December 15, 1997.