Data Analysis Project 1 - Report

Taylor Grimm, Frank Shen, Yurong Chen, Nate Byford

February 2022

1 Introduction

This study is about how the rate of oxygen consumption of crabs varies when considering three different species, three pre-determined temperature levels, and sex as factors. The data are 72 records of the oxygen consumption rate across these three factors: species (one/two/three), sex (male/female), and temperature levels (low/med/high). The rate of oxygen consumption is a continuous variable. Species, sex, and temperature levels are categorical variables. The exploratory analysis and a multiple regression analysis reveal that temperature levels and species have evident effects on the oxygen consumption of crabs and that temperature levels provide the most influence. There are also interactions between species and temperature levels and between species and sex so that effects of species on the rate of oxygen consumption are different for different genders or temperature levels.

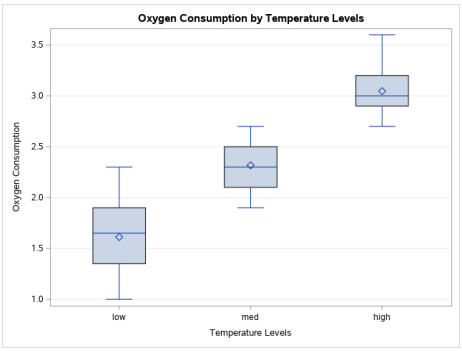
Our report is organized as follows: we first discuss exploratory analysis results based on some data visualized plots. Next, we describe our analysis method and discuss the results. The final part is our conclusion and further discussion.

2 Exploratory Data Analysis

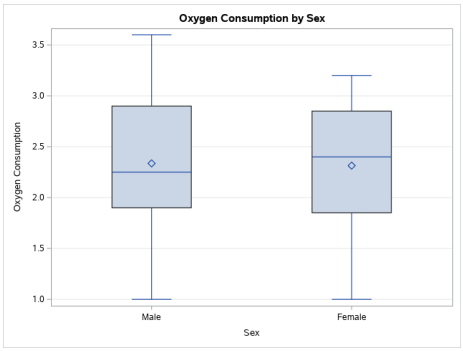
To get a better idea of the effect different variables have on the data, we made box plots for each variable. Looking at oxygen consumption by temperature levels in Figure 1a, we see that there appears to be a sizable difference in the means and medians for the temperature level groups for oxygen consumption. Crabs with the lowest rate of oxygen consumption are in the low temperature group and crabs with the highest rate of consumption are in the high temperature group.

Next, we looked at the effect of species on oxygen consumption in Figure 1c. From the plot there appears to be no significant difference in mean oxygen consumption based on the species of crab. The last variable to looked at was the sex of the crabs. In Figure 1b we observe that the mean oxygen consumption does not seem to vary much by sex. But, there appears to be more variability of oxygen consumption rates in male crabs compared to female crabs.

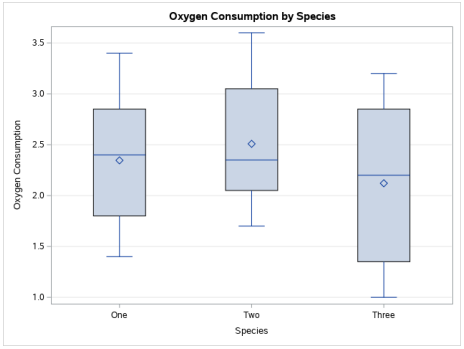
Then, we used some grouped box plots to explore any potential interactions among three factors. In Figure 3, we observed Species Two behaves differently from the other two species. Male crabs of Species One and Three tend to consume less oxygen compared to female crabs. However, in contrast, male crabs of Species Two consume more oxygen than female ones. We also noticed in Figure 3 crabs of Species Three are more sensitive to changes in temperature levels than the other two species. When the temperature drops, the rate of oxygen consumption of crabs of Species Three decreases much faster than crabs of Species One and Two.



(a) Observed rate of oxygen consumption for crabs at each temperature level.



(b) Observed rate of oxygen consumption for crabs by sex.



(c) Observed rate of oxygen consumption for crabs by species.

Figure 1: Boxplots of the observed rate of oxygen consumption for each temperature level, sex, and species.

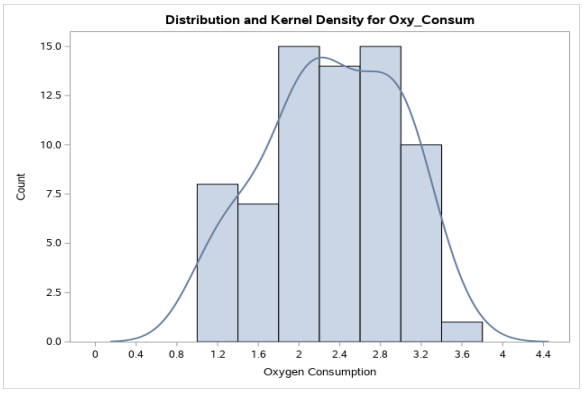


Figure 2: Histogram of the observed rates of oxygen consumption for each crab. A kernel density estimate is also overlaid with a line.

Another way we explored the data was to use a Kernel Density Estimate (KDE) to get an idea of the distribution of oxygen consumption. In Figure 2, the KDE of oxygen consumption is shown over a histogram of oxygen consumption. Here we can identify the distribution as being symmetric and fairly close to normal.

3 Methods

To understand and estimate potentially significant effects of sex, species, and temperature on the rate of oxygen consumption of crabs, we used an analysis of variance (ANOVA) test with the rate of oxygen consumption as the response variable and included the main effects of sex, species, and temperature, along with interactions between each effect, as explanatory variables. ANOVA is well-suited to understanding the relationship between species, sex, and three pre-determined temperature levels on the rate of oxygen consumption of crabs because it provides us with p-values associated with each variable, allowing us to conclude which variables have a statistically significant effect on the rate of oxygen consumption of crabs.

The linear model is as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{y}$ is a length 72 vector representing the observed values of the rate of oxygen consumption, $\boldsymbol{X}$ is a 72×48 matrix containing a column of 1's and columns of indicator variables denoting the sex, temperature level, species, and two-way and three-way interaction terms between each variable corresponding to each of the 72 observations, with each observation corresponding to a row of $\boldsymbol{X}$. $\boldsymbol{\beta}$ is a length 48 vector containing an intercept term and the effects of sex (male/female), temperature (low/med/high), species (one/two/three), and all two and three-way interactions on the rate of oxygen consumption, and $\boldsymbol{\epsilon}$ is a length 72 vector representing the error term for each observation.

Alternatively, the model can be written as follows:

$$y_{ijkt} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkt}, \tag{2}$$

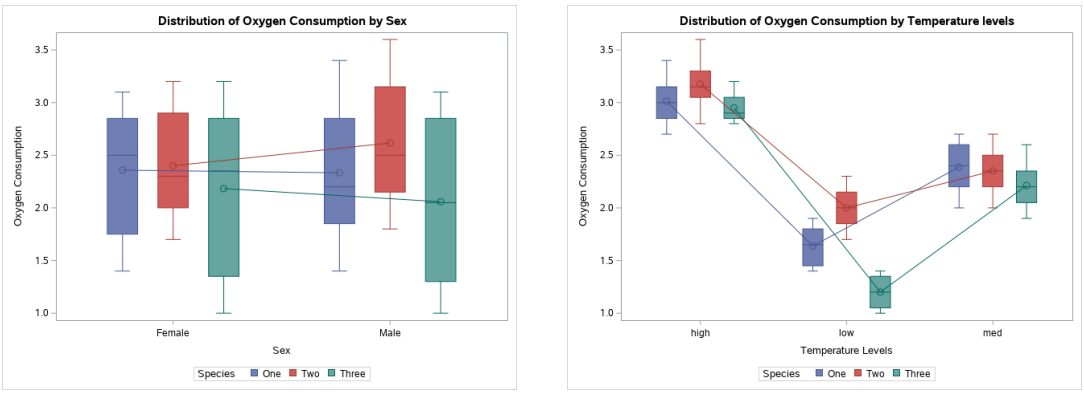


Figure 3: Boxplots of the observed rate of oxygen consumption for each sex grouped by species, and each temperature level grouped by species

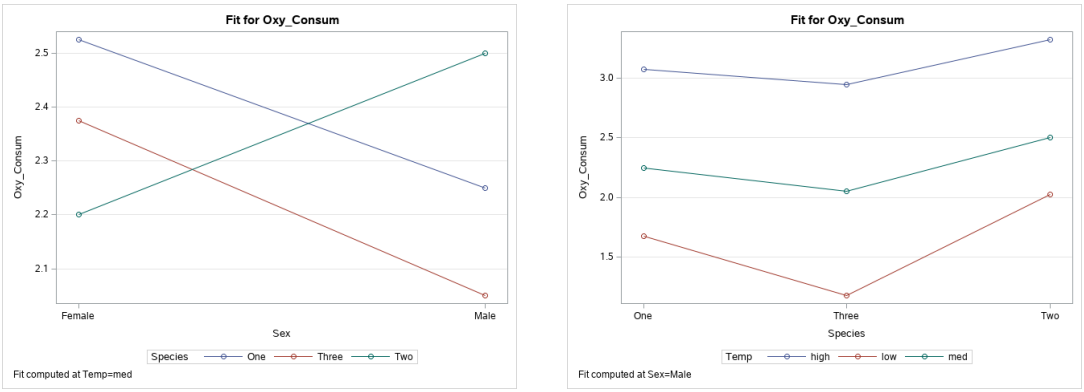


Figure 4: Plots illustrating interactions in the model. The plot on the left illustrates the interaction between sex and temperature while the plot on the right depicts the interaction between species and temperate. Both plots include the rate of oxygen consumption on the y-axis.

where y_{ijkt} represents the observed rate of oxygen consumption for crab t at levels i, j, and k of each variable, μ is the grand mean, α_i represents the main effect of species $i \in \{\text{one, two, three}\}$, β_j represents the main effect of temperature level $j \in \{\text{low, med, high}\}$, γ_k represents the main effect of sex $k \in \{\text{male, female}\}$, $(\alpha\beta)_{ij}, (\alpha\gamma)_{ik}, (\beta\gamma)_{jk}$, and $(\alpha\beta\gamma)_{ijk}$ represent the interaction effects of each respective variable at the given level of that variable, and ϵ_{ijkt} represents the error term for crab t at factor levels i, j, and k.

Our analysis of the crab data was done using PROC GLM in SAS software. This procedure provides analysis of variance (ANOVA) sums of squares output in addition to estimated regression coefficients, standard errors, p-values, and diagnostic information. Results from ANOVA and multiple regression can only be properly considered after verifying that the following assumptions have been met: a linear relationship between the response and any quantitative explanatory variables, independence of observations, normality of the response variable, and homoscedasticity.

Since each explanatory variable for this dataset is categorical, we do not need to check if there is a linear relationship between each explanatory variable and the response. We also assume independence between measurements for each observation, which seems reasonable in this situation. As discussed in section 2, Figure 2 shows a histogram along with a kernel density estimate of the measured rate of oxygen consumption for each crab. The data appears to be fairly symmetric, and we conclude that the normality assumption is reasonable. Lastly, we see diagnostic plots in Figure 5 for the final model fit produced by SAS software, and the plots in the top-left and top-center show that the residuals have no pattern and appear to be randomly and evenly spread across 0, indicating homogeneity of variance. Therefore, the assumptions required for ANOVA have been satisfied.

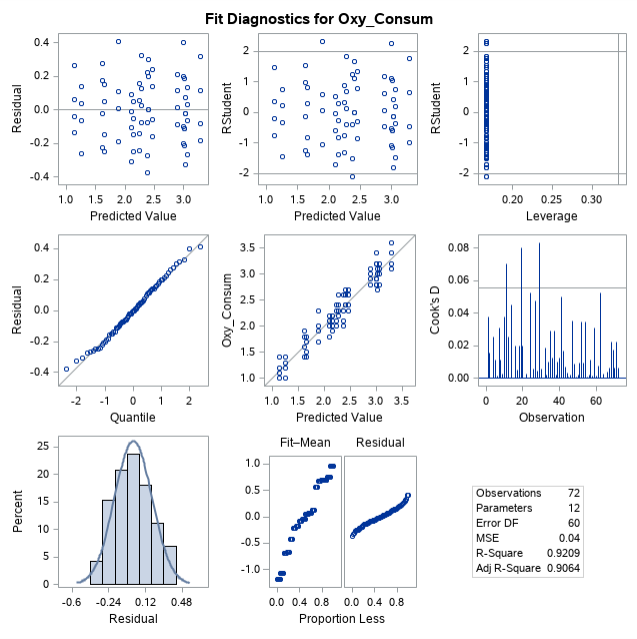


Figure 5: Diagnostic panel produced by SAS software after running PROC GLM with rate of oxygen consumption (Oxy_Consum) as the response variable and species, sex, temperature, and the interaction between species and sex and species and temperature as explanatory variables.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	28.35000000	1.66764706	44.91	<.0001
Error	54	2.00500000	0.03712963		
Corrected Total	71	30.35500000			

R-Square	Coeff Var	Root MSE	Oxy_Consum Mean
0.933948	8.287764	0.192691	2.325000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Species	2	1.81750000	0.90875000	24.48	<.0001
Temp	2	24.65583333	12.32791667	332.02	<.0001
Species*Temp	4	1.10166667	0.27541667	7.42	<.0001
Sex	1	0.00888889	0.00888889	0.24	0.6266
Species*Sex	2	0.37027778	0.18513889	4.99	0.0103
Temp*Sex	2	0.17527778	0.08763889	2.36	0.1041
Species*Temp*Sex	4	0.22055556	0.05513889	1.49	0.2196

Table 3: Tables containing type 1 sums of squares, p-values, and other statistics associated with the full model including the main effects of Species, Temp, and Sex, along with interactions between all terms.

4 Results and Discussion

After using PROC GLM to fit the full ANOVA model to the data, we found that the R-Square value is 0.934, which is quite high, so our regression model explains the variance in the rate of oxygen consumption pretty well. Also, the Root MSE is 0.19, which indicates that our model predicts the oxygen consumption rate relatively accurately when compared to the scale of the observed oxygen rates.

When we look at the p-value for species and temperature in Table 3, we see that the p-value is less than 0.0001. Therefore we can conclude that species and temperature both have significant effects on the rate of oxygen consumption of crabs. However since the p-value for sex is 0.6266, we conclude that sex does not have a significant influence on crabs' rate of oxygen consumption. Furthermore, according to Table 3, the p-values for the interactions between species and temperature levels and between species and sex are less than 0.05, so we can say that the rate of oxygen consumption of crabs is affected not only by species, sex, or temperature, but by combinations of species and sex and temperature and species.

In Figure 4, we can see plots that show the interaction between these variables. In the plot on the left, we see that the lines are not parallel, which shows us the interaction between sex and species. In particular, we see that the rate of oxygen consumption for crabs decreases for males compared to females in species one and three while it increases for males in species two. Additionally, we see in the plot on the right that, at a

low temperature, species three appears to have a much lower rate of oxygen consumption than species one, while species two has a much higher rate of oxygen consumption. The differences in these rates appears to be different for the different species at a low temperature compared to at a medium or high temperature.

5 Conclusion

In the future, it would be useful to conduct post-hoc tests in order to specifically quantify and identify which factor levels lead to an increased or decreased rate of oxygen consumption for crabs. However, through our ANOVA model, we were able to identify statistically significant effects of species, temperature, and the interactions between species and temperature and species and sex. These variables are what should be investigated more in the future in order to better understand factors affecting the rate of oxygen consumption in these species of crabs.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	27.95416667	2.54128788	63.51	<.0001
Error	60	2.40083333	0.04001389		
Corrected Total	71	30.35500000			

R-Square	Coeff Var	Root MSE	Oxy_Consum Mean
0.920908	8.603644	0.200035	2.325000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Species	2	1.81750000	0.90875000	22.71	<.0001
Temp	2	24.65583333	12.32791667	308.09	<.0001
Sex	1	0.00888889	0.00888889	0.22	0.6391
Species*Temp	4	1.10166667	0.27541667	6.88	0.0001
Species*Sex	2	0.37027778	0.18513889	4.63	0.0135

Table 7: Tables containing type 1 sums of squares, p-values, and other statistics associated with the final model including the main effects of Species, Temp, and Sex, along with the interaction between Species and Temp and the interaction between Species and Sex.

6 Appendix - SAS Code

```
/************************
Data analysis Project 1

EDA Plots
************************/

*ods graphics / reset width=6.4in height=4.8in imagemap;

options center nodate pagesize=80 ls=70 nolabel;

/* Simplified LaTeX output that uses plain LaTeX tables  */
ods latex path='/home/u59151287/LaTeX/Data_Analysis_Project1'
 file='data_analysis_project1.tex' /*style=journal*/
stylesheet="sas.sty"(url="sas");


* Change your file path below as needed;
proc import file='/home/u59151287/Data Folder/Project1.csv'
out = crab_dat
dbms = csv
replace;
```

```
run;

* Reformat data;

proc format;
/*      value $Tem */
/*              'high' = 'High' */
/*              'med' = 'Median' */
/*              'low' = 'Low'; */
        value $Gd
                'M' = 'Male'
                'F' = 'Female';
        value $Sp
                'one' = 'One'
                'two' = 'Two'
                'thr' = 'Three';
run;


data crab_dat;
set crab_dat;
/* format Temp $ Tem. */
format
           Sex $ Gd.
           Species $ Sp.;
run;

* Box plots;
title 'Oxygen Consumption by Temperature Levels';
proc sgplot data=crab_dat;
        vbox Oxy_Consum / category=Temp;
        yaxis grid;
        xaxis discreteorder=data;
        label Oxy_Consum = 'Oxygen Consumption' Temp = 'Temperature Levels';
run;


title 'Oxygen Consumption by Species';
proc sgplot data=crab_dat;
        vbox Oxy_Consum / category=Species;
        yaxis grid;
        xaxis discreteorder=data;
        label Oxy_Consum = 'Oxygen Consumption';
run;

title 'Oxygen Consumption by Sex';
proc sgplot data=crab_dat;
        vbox Oxy_Consum / category=Sex;
        yaxis grid;
        xaxis discreteorder=data;
```

```
        label Oxy_Consum = 'Oxygen Consumption';
run;

* Comparative Boxplots;

title 'Distribution of Oxygen Consumption by Sex';
proc sgplot data=crab_dat;
   vbox Oxy_Consum / category=Sex group=Species connect=mean;
   xaxis label="Sex";
   keylegend / title="Species";
   label Oxy_Consum = 'Oxygen Consumption';
run;

title 'Distribution of Oxygen Consumption by Temperature levels';
proc sgplot data=crab_dat;
   vbox Oxy_Consum / category=Temp group=Species connect=mean;
   xaxis label="Temperature Levels";
   keylegend / title="Species";
   label Oxy_Consum = 'Oxygen Consumption';
run;

* Distribution of the  rate  of  oxygen consumption;
proc kde data=crab_dat;
        univar Oxy_Consum(bwm=0.4) Oxy_Consum(bwm=0.6);
        label Oxy_Consum = 'Oxygen Consumption';
run;

* Model with no interactions;
proc glm data=crab_dat alpha = 0.05 plots = (diagnostics residuals);
class Species Temp Sex;
model Oxy_Consum=Species Temp Sex / solution;
run;

* Full model with all interactions;
proc glm data=crab_dat alpha = 0.05 plots = (diagnostics residuals);
class Species Temp Sex;
model Oxy_Consum=Species  Temp  Sex / solution;
store full_glm_model;
run;

* Create interaction plots;
title 'Interaction Plot for Species and Temperature';
proc plm restore=full_glm_model;
effectplot interaction(x = Species);
label Oxy_Consum = 'Oxygen Consumption' Temp = 'Temperature Level';
run;

title 'Interaction Plot for Species and Sex';
proc plm restore=full_glm_model;
effectplot interaction(x = Sex);
```

```
label Oxy_Consum = 'Oxygen Consumption';
run;
title;


* Final model;
proc glm data=crab_dat alpha = 0.05 plots = (diagnostics residuals);
class Species Temp Sex;
model Oxy_Consum= Species Temp Sex Species*Temp Species*Sex/ solution;
store final_glm_model;
run;



ods latex close;
quit;
```