Mathematical Statistics Project

Dillon Allen

Problem 1

In this problem, we will investigate the hypothesis of whether the mean time fish spend guarding their eggs is equal to the mean time they spend fanning. Without blocking the fish by pH or Gender, we have the t.test value of 8.6491 and a binomial test p-value of 2.266e-7, meaning we reject the null hypothesis of the mean times being equal. Breaking the fish up into pH groups, we can look at the scatter plots to get an idea of their behavior based of their pH.

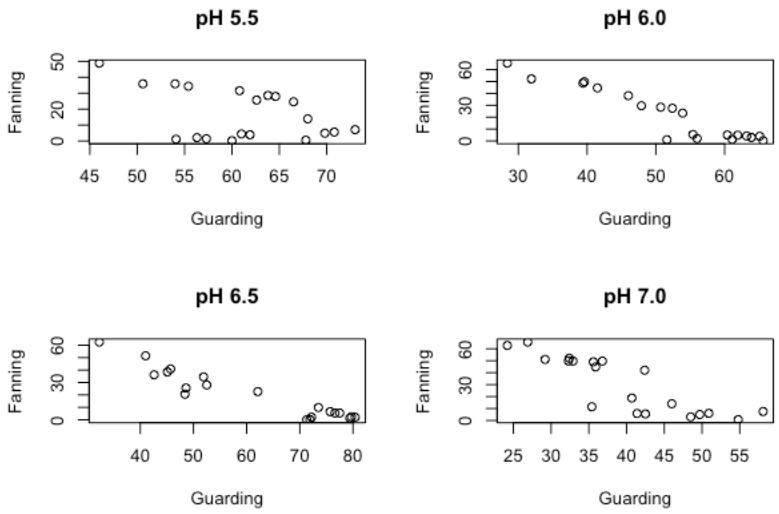

Figure 1: Fanning times vs Guarding times by pH level

The plots here tell us that for pH 5.5, 6.0 and 6.5 there seems to be more concentration on the higher guarding times rather than the time spent fanning. For the pH of 7.0, we can see that the trend is still moving in the same direction, but the grouping of points is more towards fanning rather than guarding. For the pH of 5.5, we get a t-test value of t = 9.8986 and sign test p-value of 4.005e-5, rejecting our null hypothesis. Similarly, for pH of 6.0 and 6.5 we get t-scores and sign test values of (4.15,0.04139) and (5.386, 0.0004025) respectively, also leading us to reject the null hypothesis. The t-score for pH 7.0 is t = 1.4236 sign test p-value of .8238. This means we cannot reject the null hypothesis for this group. Splitting the fish by genders, we get the following scatterplot

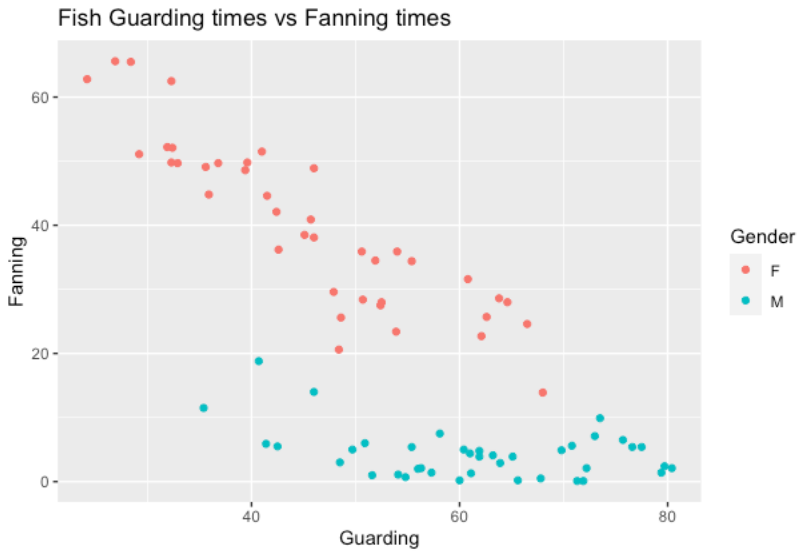

Figure 2: Fanning times vs Guarding times, colored by Gender. (Experimenting with ggplot2)

The males tend to spend most of their time guarding the eggs rather than fanning, where the females have a distributed amount of time between fanning and guarding. The paired t-test and sign tests for the male group of fish came out to (26.289, 1.89e-12), leading us to reject the null hypothesis yet again. For the females, we see the opposite, where the t-score and sign test resulted in values of (1.4853, 0.4296), which indicates that there is not statistically strong enough evidence to reject the null hypothesis.

Problem 2

In this problem, we were given a generic dataset with seven possible predictor variables and a single response variable y, with the task of fitting a regression line. Since the response variable histogram indicated a skewed-right distribution, we used log(y) as the transformation to transform the response data. After doing so, we fit a comprehensive model

$$y = -0.01234x_2^2 + 1.63523x_3 + 1.92263x_5 + 0.93159x_7 + 3.59966$$

With an $R^2 = 0.7809$. Judging by the residual vs predictor variable scatter plots, it looked like x_2 was the variable that would benefit the model best by squaring. Although this may not be the best model, this was the one I created that performed the best after a few hours of generating different transformations and equations.

Problem 3

In this problem, we looked at the use of classification trees when it comes to predicting glaucoma. After fitting our model, our confusion matrix came out to

	Predicted	
Observed	Normal	Glaucoma
Normal	81	17
Glaucoma	6	92

For just using rpart to fit the classifier, I believe this model works pretty well. An application for this type of classification can be for target discrimination in military applications or even used for exoplanet detection given a range of parameters. With huge exoplanet surveys currently being conducted, a lot of candidate data is being produced without an efficient way to categorize what may be worthy of further investigation versus a false detection. Using a classifier can help narrow down the list of candidates while leaving room for the physicist or data scientist to create more sophisticated models to work on the filtered results.

Problem 4

The country I chose to explore was Spain, where there were some unique findings. The initial reason I investigated Spain was because life expectancy did not drop nearly as much as the other major countries during World War II. Investigating the history of Spain, the ruling regime at the time, the Francoists, stayed neutral during the war. Due to this, there was little damage done to the civilian population and life expectancy during that time period. On the other hand, there was a dip in life expectancy during the 1935-1940 era due to the civil war due to the military coup. According to Gapminder, there was only really a strong dip during the 1917s (Spanish flu/WW1) and some turbulence in the life expectancy during the civil war. I question the accuracy of some results, as there were many reports of reprisals, concentration camps and mass killings during the 1940s which saw a meteoric rise in life expectancy vs income. Looking at the Life Expectancy vs Population option, we also see the same meteoric rise in life expectancy but with very little variation in the population size, although reports of deaths in the 100,000s were reported during the early reign of the Francoist regime.