Exploratory Data Analysis: Finches' Dataset

1. Introduction

In the Finches' dataset we can find 100 records of *Geospiza fortis* that survived or not a drought in 1977. In this paper, I will explore the data available to find some intuition on what could have determined the survival of those individuals, as well as some other relationships that might become apparent.

---

2. Preparing the data

The dataset provided had some formatting mistakes, such as misplaced and missing tabulations, which had to be fixed before use.

The variables for each of the 100 individuals are the following:

Qualitative	Quantitative, discrete	Quantitative, continuous
1. Species 2. Sex	1. Band 2. First adult year 3. Last year	1. Weight 2. Wing 3. Tarsus 4. Beak length 5. Beak depth 6. Beak width

What each of the variables represent is described here. Since Band is an identifying arbitrary value and Species is always the same one, I consider those variables useless for the analysis.

From the available data, we can compute some more variables to see if those are important to understand the data.

Age can be computed as (Last year - First year + 1). It is a quantitative, discrete variable.

Survival is True if the Last year is greater than 1977. It is a qualitative/logical variable.

---

3. Understanding the data

Some basic statistics of the available variables are shown in Appendix 1. For continuous variables, I plotted the frequency of the observed measures, I did a box plot and run a normality test to continue the analysis.

So far, the analysis reveals some remarkable trends in our data:

- In this dataset, males are overrepresented compared to females.
- For sex, missing values are more than a third of the entire sample, which could be detrimental to the power of statistical tests.
- First adult year confirms that all the individuals were measured first before the drought.
- As described in the information provided with the dataset, the quantity of individuals which survived and which didn't is the same, so the dataset is balanced in this sense.
- The dataset is not balanced for all the groups that we could consider in the population. The age of the individuals shows a peak at 2-3 years. The 2 year old individuals may be the ones that died during the drought.
- Since the age of an organism and its sex influence some body measures, the unbalance of sex and age can result in those other variables being unbalanced.
- Weight and wing length are not normally distributed, while beaks measures and tarsus are normally distributed.

3.1 Basic relationships between variables

A correlation matrix was constructed, representing correlation between continuous variables, box plots and histograms for categorical variables. It is available in Appendix 2.

The scatter plots representing body measures (weight, wing, tarsus and beak dimensions) show the most obvious trends in the data. Their correlations range from 0.461 (tarsus ~ wing) to 0.877 (beak width ~ beak depth).

I expected the age to present a strong correlation to body measurements, but it does not seem so. However, box plots from different sex and survival groups apparently show differences between them. For example, wing lengths seem to be larger in male and survival groups vs female and non-survival groups. These hypothesis need to be statistically tested in the following sections.

---

4. Hypothesis testing

4.1 Grouping

In this section, I will assess two main questions:

1. Is there sexual dimorphism in this species for the body traits present in the dataset?
2. Do body measures relate to survival during the drought?

These questions need to be formulated in a statistically valid way, i.e., is there a significant difference in body measures between groups (such as male/female, survival/not survival)?

4.2 Test choice

Since the groups above stated are independent from in other (an individual from one group cannot be present in the other one), we can consider the (sub)samples for the tests to be independent.

As previous tests confirm, most of the quantitative variables of the dataset follow a normal distribution. Thus, an t-test for independent samples could be performed to assess differences in their means. However, we still need to check if the variance of the groups to be compared are the same or not. All of this is considered in the code, which will pass the corresponding parameter to the test (see Appendix: code).

For the non-normal quantitative variables (weight and wing), I still studied if the distribution of each of the groups was normal. The result (not shown) was that wing is normally distributed in each of the groups, while weight is not. Due to this, I did not performed a t-test for weight, but a Mann-Whitney U test, which is non-parametric.

4.3 Results

The complete detailed report of the results if available as Appendix 3.

In all the cases, there was homogeneity in the variance, so the regular t-test and Mann-Whitney tests could be performed.

The following table shows the significance of the differences found:

	Weight	Wing	Tarsus	Beak length	Beak depth	Beak width
Sex	*0.09156*	*0.001952*	*0.0916*	*0.1888*	*0.1171*	*0.3949*
Survival	*0.0002102*	*0.001482*	*0.07186*	*0.0004474*	*0.0015*	*0.01136*

From the table above we can observe that there is sexual dimorphism in the wing size, but not for the rest of the body measures.

Beak length is the body measure that is the most distinct between the individuals that survived and the ones that did not.

---

5. Conclusions and considerations

The dataset provided contained 34 individuals of unknown sex and only 19 females. This fact limited the statistical power of the tests performed and could have distorted the sample distribution from the actual population distribution. Here, the t-tests performed did not use those 34 records with missing values. Even though there are more advanced techniques which can use incomplete datasets for analysis, the best approach to increase the quality (reduce the possible errors in conclusion-making) is always to increase the sample size and the sampling quality.

All body measures but tarsus length were positively correlated to the survival of the individuals. This makes sense in the context of the natural selection and selective pressure caused by the drought. The results show that there is indeed an association between increased beak and wing sizes and survival.

We could further hypothesize that this differences in beak dimensions would facilitate finding food and other resources to an individual, or even that the increased wing size would help them reaching areas not affected by the drought. Nevertheless, this assumptions should not be made without the required knowledge in bird biology, so I will avoid making such statements.