Doggy Demographics:
Discerning Dimensions of Depth, Dye, and Dartiness

Collin Nolte

Introduction

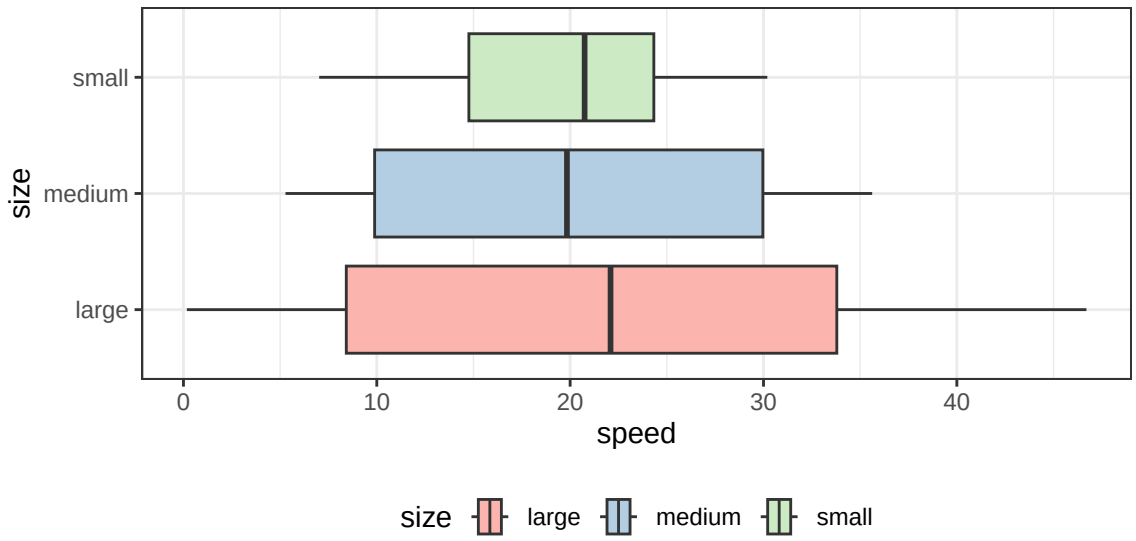
My first paragraph will be an introduction to my report, including broad-stroke information about my study and my goals. In for this particular study, there are primarily two research questions that I have. The first asks if the size and color of dogs are independent, while the second asks if dogs differ in how fast they are, based on their size. Notice how my report starts with regular text, too – there are no messages, warnings, or R code that appear anywhere.

Next, I will describe the data I have and how it was collected. For example, for this study I visited several parks, befriending various canines until I had collected data on 400 individual observations. For each dog, I recorded the dog's breed, its size, color, and the highest speed obtained in a one hundred meter dash, measured in miles per hour. As one of my primary research questions relates to size and color, I have included summary information below. I don't have to include formatted summary tables, but if I wanted to, the Rmd file associated with this document shows how to create this one (included on the course webpage).

Table 1: Table with size and color

	black	brown	white	yellow
large	100	0	25	75
medium	50	25	25	0
small	50	25	0	25

As the relationship between size and speed makes up our second research question, I have presented a plot demonstrating this association below.



Methods

My first research question addresses the issue of independence between the size of a dog and its color. As these are both qualitative variables, I will be conducting a χ^2 test of independence at a significance level of $\alpha = 0.05$. I should note that if I try and copy α or χ^2 directly into my document, I will *not* be able to get it to knit. Instead, I should look at the Rmd file to see how these symbols are created using Latex.

My second research question asks if there is any difference in speed between the different sizes of dogs. As this question involves both a categorical variable and a quantitative variable, I have a number of options available. First, I could do a simple ANOVA to test the hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3.$$

Alternatively, I could could use linear regression to see if there is an association or not. Because I am a man of rather refined taste, I have decided on using ANOVA, again testing at the $\alpha = 0.05$ level.

Results

Generally speaking, there will seldom be an occasion in which it is appropriate to include raw R output, including output from statistical tests. For example, our first test is a χ^2 test of independence. I can see the Rmd file that I have conducted this test *without* printing the output into R. From this test, I should write out the results directly.

In testing our first hypothesis, that is, the independence of size and color of our dogs, we derive a χ^2 test statistic of $\chi^2 = 112$ with 6 degrees of freedom, resulting in a p-value of $p < 0.0002$. As $p < \alpha = 0.05$, we handedly reject the null hypothesis that size and color are independent.

Next we consider an ANOVA model to investigate whether or not there are differences in group means in speed between dogs of different sizes.

Our ANOVA results in an F statistic of $F = 1.9$ with 2 and 397 degrees of freedom, giving us a p-value of $p = 0.15$. As $p > \alpha$, we fail to reject our hypothesis that dogs of each size have the same average speed. Note further that it is completely acceptable to have a statistical test that does not reject our null hypothesis.

Discussion

We started this study in pursuit of two questions: is the size and color of a dog independent, and is there evidence to suggest that dogs differ in their average speed, given their size. Based on our statistical tests, we have concluded that there does appear to be an association between size and color, while there is no evidence of any relationship between size and speed. Both of these outcomes are surprising in that both are contrary to what we expected to be true prior to beginning our study. Although there are no statistical tests to confirm this, these results appear to suggest that perhaps this data has been artificially generated, lending support to the theory that we do indeed live in a simulation.