Test for Association Between a Categorical and Quantitative Variable

Thomas Scofield

November 15, 2021

Idea of ANOVA

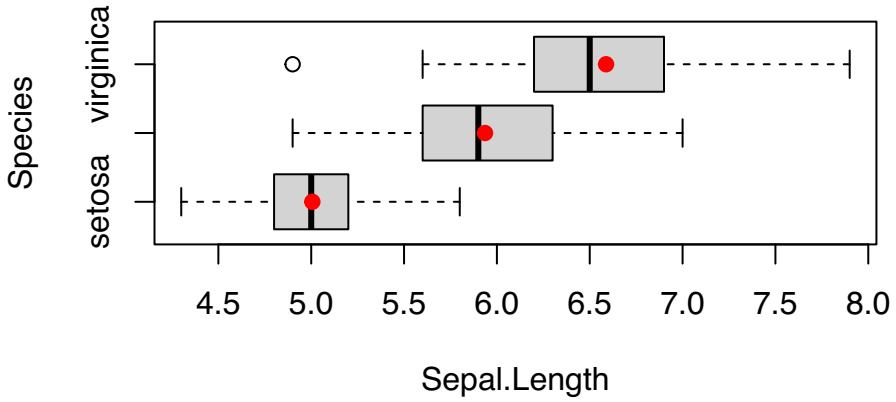
Consider the `iris` data set, for which a partial data table is given above. When we look `Sepal.Length` broken down by `Species`, we see that the samples have different means:

```
irisMeans <- mean(Sepal.Length ~ Species, data=iris)
irisMeans
```

```
##     setosa versicolor  virginica
##      5.006      5.936      6.588
```

Different sample means do not always lead to the conclusion that there are different population means. We might look at side-by-side boxplots to assist our intuition about whether the population means are different. What follows is a standard boxplot (the lines through the interior of the boxes give the locations of the three sample medians) enhanced by the incluion of a solid red dot at the locations of the group means.

```
boxplot(Sepal.Length ~ Species, data=iris, horizontal=TRUE)
points(c(1,2,3) ~ value(irisMeans), pch=19, col="red")
```

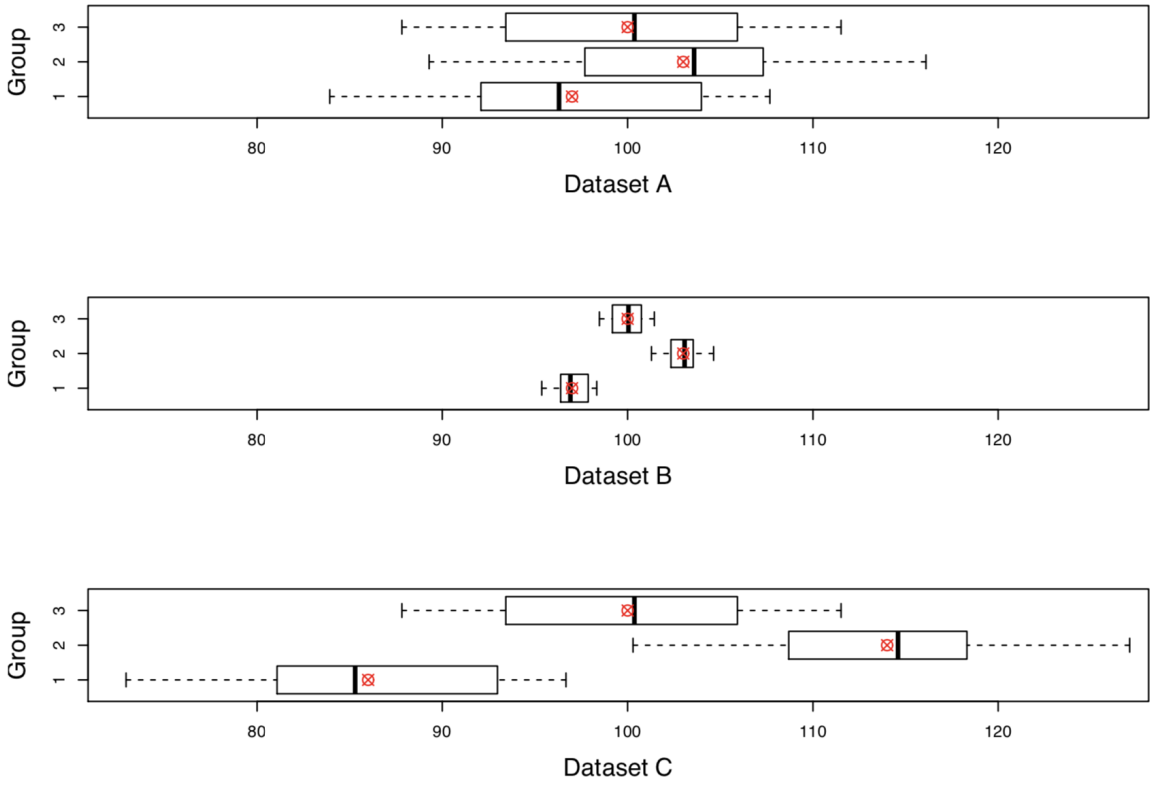


How does this plot add or detract from the evidence that the population means are not all the same across the three species? What is our basis for comparison? Consider this figure from the Lock 5 text. In particular, start by focusing on Dataset A, the first set of side-by-side boxplots. As with my boxplots above, there are boxplots for each of three "species", enhanced with a red mark is placed at the location of the sample means. In truth, this is made-up data to illustrate a point. We see the three sample means are not all the same (not vertically aligned), but are they far enough apart to provide convincing evidence that the corresponding population means are not all the same?

But compare with the three boxplots of Dataset B. Those have means in the exact same locations as the three means of Dataset A, yet somehow, it seems like the boxplots for Dataset B are more convincing of a difference in underlying population means than the boxplots for Dataset A. Why? Because the boxplots

rfor the three species in Dataset A are more spread out, overlapping more than occurs with the boxplots for Dataset B.

Now compare the three boxplots for Dataset A with those for Dataset C. The boxplot for Group 1 in Dataset C has exactly the same *range* (length between ends of the whiskers) and *IQR* (width of box) as for Group 1 in Dataset A. The same sort of consistency in range and IQR has been maintained for Groups 2 and 3 (only speaking about Datasets A and C here). Yet it seems that the boxplots for Dataset C are more convincing as evidence that group population means are different than the boxplots of A.



In this module we consider bivariate data where the explanatory variable is categorical with k distinct values (but, unlike the 2-sample mean tests of Chapter 6, we do not require $k = 2$), and the response variable is quantitative. For the iris data, the categorical variable is `Species` and has three values (corresponding to the three boxplots), and the response variable is `Sepal.Length`. Our main focus is a test with null hypothesis that all population means are equal

$$\mathbf{H}_0 : \ \mu_1 = \mu_2 = \cdots = \mu_k,$$

and the alternative hypothesis is that at least one of the population means is different from another.

Given the intuition arising from the side-by-side boxplots above, we must

- have a dataset which is comprised of independent random samples from each group,
- produce a test statistic that takes into account both
- how different the individual means are, and
- the spread of individual samples.

This is the idea of 1-way ANOVA (ANalysis Of VAriance).

The F-statistic

First, some symbol definitions. In our dataset we have k different groups/species/populations, and the categorical explanatory variable identifies the group to which each case belongs. We can take group sample means of the quatitative response fariable, means we refer to as $\overline{x}_i$. That is, $\overline{x}_1$ is the average response value for cases in Group 1, $\overline{x}_2$ is the average response value for cases in Group 2, and so on. In R, a command like

```
mean( responseVar ~ explanatoryVar, data=dataset )
```

would give us $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k$ all at once.

We could also treat the dataset as a single whole, not distinguishing between individual groups/populations, and calculate the grand mean $\overline{x}$. The corresponding command might be something like

```
mean( ~ responseVar, data=dataset )
```

Along with symbols $\overline{x}, \overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k$ we use n_i to denote the size of the sample from Group i, and s_i to denote the standard deviation of the sample from Group i. (This sort of notation was employed in 2-sample mean problems found in Section 6.12 of the Lock 5 text.)

After viewing some side-by-side boxplots in the previous section, we stated some criteria for a test statistic. The quantities we consider next are for the purpose of constructing a test statistic that meets these criteria. First, we have the quantity SSG that helps to measure *between-group-variability*:

$$SSG \;=\; \sum n_i(\overline{x}_i - \overline{x})^2.$$

Here the SS stands for "sum of squares". Another quantity, SSE,

$$SSE \;=\; \sum (x - \overline{x}_i)^2,$$

helps to measure *within-group-variability*. Yet another sum of squares, SST,

$$SST \;=\; \sum (x - \overline{x})^2,$$

helps to measure *total variability*.

Example. Say the sampled values for Groups 1, 2 and 3 are as follows:

- Group 1: 15, 18, 17
- Group 2: 14, 11, 13
- Group 3: 16, 17, 19, 17

Starting with the grand mean, its value is

$$\overline{x} \;=\; \frac{1}{10}(15+18+17+14+11+13+16+17+19+17) \;=\; 15.7.$$

Next, the mean $\overline{x}_1$ for Group 1 is

$$\overline{x}_1 \;=\; \frac{1}{3}(15+18+17) \;=\; 16.667,$$

with the result that Group 1 contributes

$$(15-16.667)^2+(18-16.667)^2+(17-16.667)^2 = 4.667 \text{ to } SSE \qquad \text{and} \qquad 3(16.667-15.7)^2 \;=\; 2.803 \text{ to } SSG.$$

Similarly, for Group 2,

$$\overline{x}_2 \;=\; \frac{1}{3}(14+11+13) \;=\; 12.667,$$

so it contributes

$(14-12.667)^2+(11-12.667)^2+(13-12.667)^2 = 4.667$ to SSE and $3(12.667-15.7)^2 = 27.603$ to SSG.

Finally, for Group 3,

$$\overline{x}_3 = \frac{1}{4}(16+17+19+17) = 17.25.$$

so Group 3 contributes

$(16-17.25)^2+(17-17.25)^2+(19-17.25)^2+(17-17.25)^2 = 4.75$ to SSE and $4(17.25-15.7)^2 = 9.61$ to SSG.

We get SSE and SSG by summing the individual contributions to each:

$$SSE = 4.667+4.667+4.75 = 14.084, \quad \text{and}$$

$$SSG = 2.803+27.603+9.61 = 40.016.$$

The total sum-of-squares is

$(15-15.7)^2+(18-15.7)^2+(17-15.7)^2+(14-15.7)^2+(11-15.7)^2+(13-15.7)^2+(16-15.7)^2+(17-15.7)^2+(19-15.7)^2+(17-15.7)$

Notice that

$$SST = 54.1 \quad \text{and} \quad SSG+SSE = 40.016+14.084 = 54.096$$

are nearly equal, and would be exactly the same but for rounding off during computation.

Our test statistic will be a ratio of between-group-variability and within-group-variability. However, a simple ratio of SSG to SSE would not quite be comparing apples to apples, so to speak. When, in Chapter 2, the Locks introduced the formula for sample *variance*, it was

$$\frac{1}{n-1}\sum(x-\overline{x})^2.$$

The variance contains a sum-of-squares, tempered (divided) by its degrees of freedom. Similarly, we consider a truer measure of variability between groups to be

$$MSG = \frac{SSG}{k-1},$$

and the measure of variability within groups to be

$$MSE = \frac{SSE}{n-k}.$$

The test statistic, called F, is the ratio

$$F = \frac{MSG}{MSE}.$$

The results of the various calculations are usually arranged in an **ANOVA table**.

Source	df	SS	MS	F-stat	P-value
Groups/Factors	$\text{df}_1 = k-1$	$SSG = \sum n_i(\overline{x}_i-\overline{x})^2$	$MSG = \frac{SSG}{k-1}$	$F = \frac{MSG}{MSE}$	P
Residuals/Errors	$\text{df}_2 = n-k$	$SSE = \sum(x-\overline{x}_i)^2$	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	$SST = \sum(x-\overline{x})^2$			