Linear Modelling

Brad Duthie

03/06/2020

Let's first make up some data and put it into a data frame. To make everything a bit more concrete, let's just imagine that we're sampling the heights of individual plants from two different species. Hence, we'll have one categorical independent variable, and one continuous dependent variable (plant height). I am just going to make up some data to work with below. The data frame below includes plant height (`height`; since this is a made up example, the units are not important, but let's make them mm) and species ID (`species_ID`). The first 10 plants (each plant is a unique row) are shown below.

height	species_ID
498.7479	species_2
373.3317	species_2
402.7320	species_1
418.8561	species_1
418.5089	species_1
440.8311	species_2
478.3306	species_2
443.0290	species_1
389.5761	species_2
351.0882	species_2

Using the linear modelling approach described by Lindeløv, the above data qualify as a simple regression with a discrete x (`species_ID`). Assuming that both species have equal variances in height, we can use a two-sample t-test in R to test the null hypothesis that the mean height of `species_1` is equal to the mean height of `species_2`. To use `t.test`, we can first create two separate vectors of heights, the first one called `species_1`.

```
species_1 <- plant_data$height[plant_data$species_ID == "species_1"];
```

Below shows `species_1`, which includes the heights of all 46 plants whose `species_ID == "species_1"`.

```
##  [1] 402.7320 418.8561 418.5089 443.0290 489.9607 439.7465 433.6579 468.6326
##  [9] 436.3861 476.7988 440.7444 433.1007 445.3568 435.0850 406.9291 449.1417
## [17] 381.2563 411.9313 389.5387 416.5154 413.7254 455.9449 404.8120 408.2896
## [25] 491.1656 431.0551 506.6516 444.6952 365.4019 404.3021 404.4872 352.3612
## [33] 414.0271 359.6948 382.7102 360.8343 420.0233 455.9284 424.1806 481.5277
## [41] 390.4399 364.6961 393.8550 394.2941 356.3389 426.2868
```

We can make a separate vector for the remaining heights for the plants of species 2 in the same way.

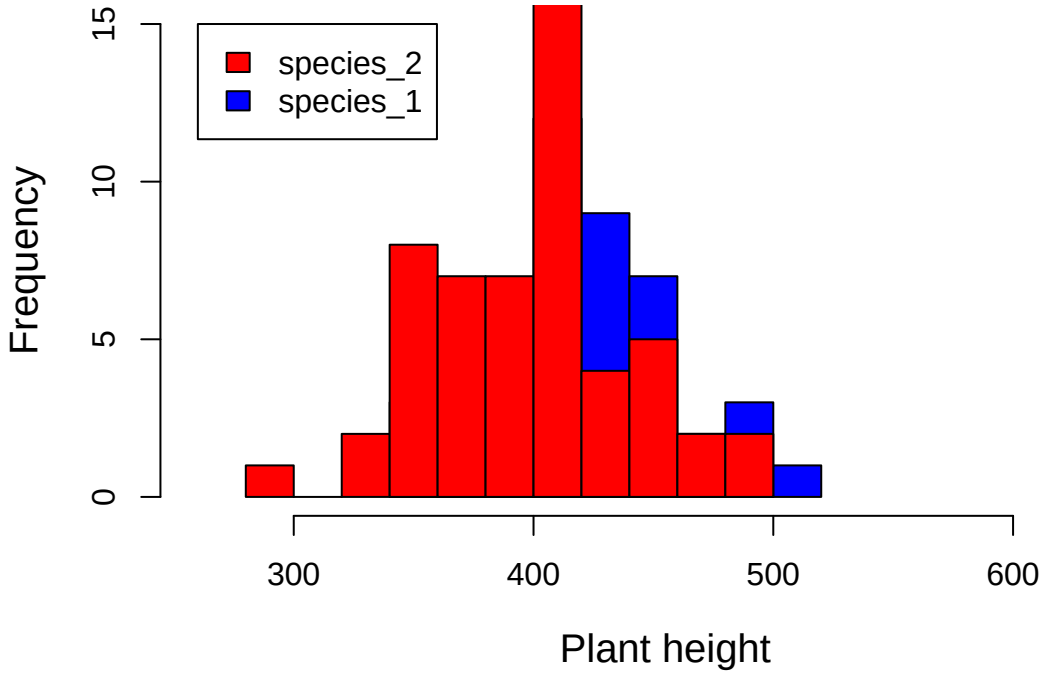
```
species_2 <- plant_data$height[plant_data$species_ID == "species_2"];
```

These 54 plant heights are shown below.

```
##  [1] 498.7479 373.3317 440.8311 478.3306 389.5761 351.0882 412.6743 452.0337
##  [9] 371.9683 455.2842 409.1347 428.7674 363.8685 365.4184 412.4938 444.4685
```

```
## [17] 447.5214 299.5538 412.1989 426.4697 409.6854 349.7132 409.0824 342.7194
## [25] 393.6626 348.7487 404.0304 394.3553 401.7397 428.2768 366.4813 343.2560
## [33] 346.8217 416.4913 418.9492 367.8585 402.4418 389.0937 377.9161 387.3182
## [41] 481.9264 468.2472 418.9458 414.3585 324.5264 401.0120 393.0444 424.6052
## [49] 416.3273 356.5363 323.9399 358.3885 418.9629 380.8652
```

It might help to plot a histogram of the two plant species heights side by side.



Visualising the histogram above, we already have a sense of whether or not knowing species ID is useful for predicting plant height. Nevertheless, with our two vectors `species_1` and `species_2`, we can run a t-test as noted by Lindeløv.

```
t.test(species_2, species_1, var.equal = TRUE);
```

```
##
##  Two Sample t-test
##
## data:  species_2 and species_1
## t = -2.757, df = 98, p-value = 0.006958
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -38.090315  -6.206244
## sample estimates:
## mean of x mean of y
##  398.4090  420.5573
```

Reading the output above, we can get the t-statistic `t = -2.7570218`. Given the null hypothesis that the mean height of `species_1` equals the mean height of `species_2`, the probability of getting such an exterme

difference between the two observed means is `p-value` < 0.006958037 (i.e., unlikely).

But this is not the only way that we can run a t-test. As Lindeløv points out, the linear model structure works just fine as well.

```
lmod1 <- lm(plant_data$height ~ 1 + plant_data$species_ID);
summary(lmod1);
```

```
##
## Call:
## lm(formula = plant_data$height ~ 1 + plant_data$species_ID)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.855 -27.556   1.034  21.033 100.339
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    420.557      5.903  71.241  < 2e-16 ***
## plant_data$species_IDspecies_2  -22.148      8.033  -2.757  0.00696 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.04 on 98 degrees of freedom
## Multiple R-squared:  0.07198,    Adjusted R-squared:  0.06251
## F-statistic: 7.601 on 1 and 98 DF,  p-value: 0.006958
```