Statistical Inference, Part 2: Basic Inferential Data Analysis

ABC

January 22, 2019

Part 2: Basic Inferential Data Analysis

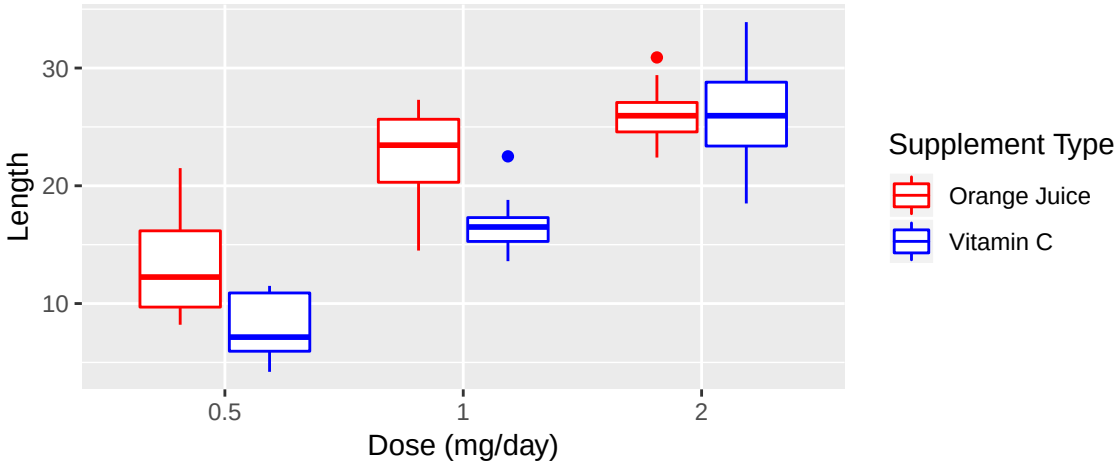
Load ToothGrowth and Perform Exploratory Analysis

The dataset is part of the datasets package. Loading `library(datasets)` will create the `data.frame` `ToothGrowth`. Because the dosage, `dose`, starts as numeric I will convert it to a factor of three levels. I prefer to work in the tidyverse, so I loaded that package and converted `ToothGrowth` to a tibble.

The dataset is fairly straightforward - it is has two experimental variables, `dose` and `supp`, the dosage in mg/day and supplement type, respectively. The length, `len`, is the observed value. First thing to do to look at it would be to plot the lengths in a set of box plots, split by the experimental variables.

Fig 4. The Effect of Vitamin C on Tooth Growth in Guinea Pigs

Odontoblast Length, by Supplement Type and Dosage



This tells us two things - for both treatments the mean growth increases with dosage within the window under examination, and in at least one situation (1 mg/day) there appears to be a statistically significant difference between the means of the different treatments.

Basic Data Summary

The basic summary of the above can be seen in the summarized mean and standard deviation, when split by dose and supplement type. Here "OJ" refers to Orange Juice and "VC" refers to Vitamin C.

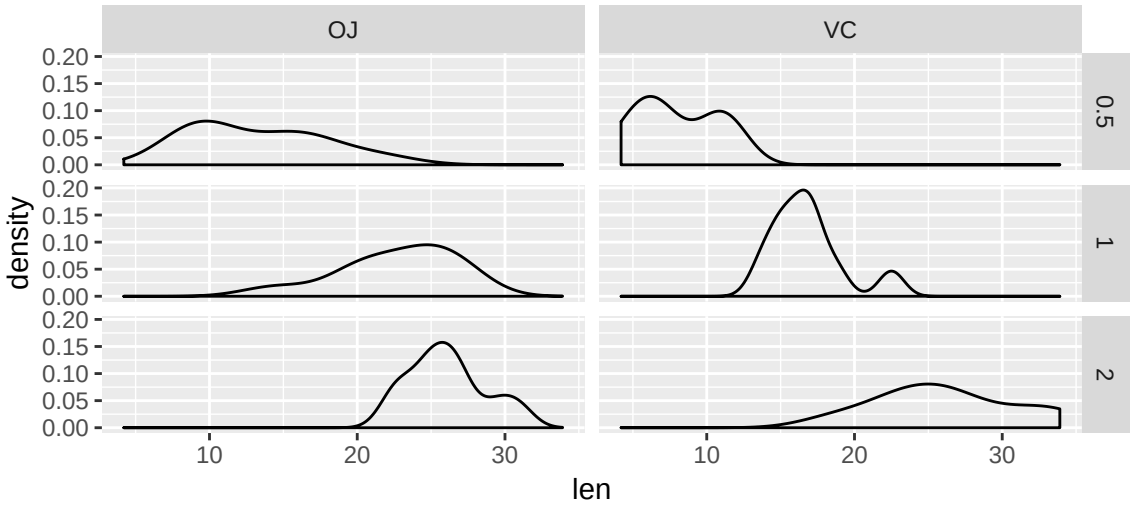
Table 1: Summary of ToothGrowth, grouped by dose and supplement

dose	supp	Average Length	Length Std. Dev.	n
0.5	OJ	13.23	4.46	10
0.5	VC	7.98	2.75	10

dose	supp	Average Length	Length Std. Dev.	n
1	OJ	22.70	3.91	10
1	VC	16.77	2.52	10
2	OJ	26.06	2.66	10
2	VC	26.14	4.8	10

We can also look at the distributions directly, although it should be kept in mind that n = 10 in each.

Fig 5. Density plots of Length, split by Dose and Supplement Type



Hypothesis Testing on Experimental Variables

One thing we can test is whether the difference of the means between the supplement type is significant at each dosage level. To do this we can set up the null hypothesis that $H_0|\mu_{OJ} = \mu_{VC}$ This can be performed as a set of two-sample t-tests. The following table shows the p-values.

Table 2: Unpaired Two-sample T-Test Between Supplements at Each Dosage

Dose	P.value
0.5	0.0064
1.0	0.0010
2.0	0.9639

The p.values indicate that for dosages 0.5 mg/day and 1.0 mg/day the distributions are such that there is a 99.4% and 99.9% chance, respectively, that the null hypothesis $\mu_{OJ} = \mu_{VC}$ can be rejected. We can not reject the null hypothesis for 2.0 mg/day.

We can also perform the t-test between the dosages to show that there is a dosage-based effect as well.

Table 3: Unpaired Two-sample T-Test Between Dosages for Each Supplement

Supplement	OJ.P.value	VC.P.value
0.5 to 1.0	0.00009	0.0000007

Supplement	OJ.P.value	VC.P.value
1.0 to 2.0	0.03920	0.0000916

All of the P-values are extremely small. The difference between 1.0 and 2.0 mg/day for OJ is the greatest P-value, although the difference is still significant ($p < 0.05$).

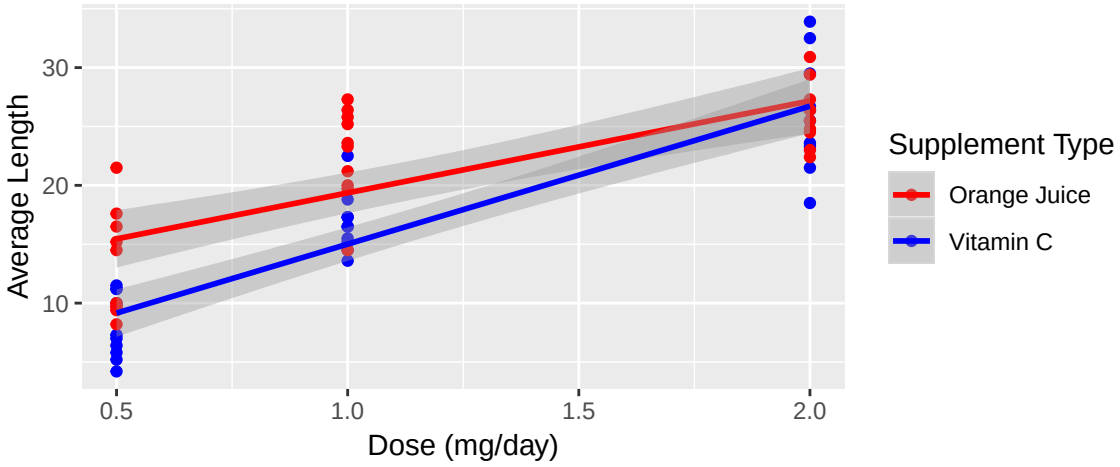
Conclusion

The data suggest that the odontoblast length is positively influenced by both orange juice and vitamin C. At concentrations of 0.5 mg/day and 1.0 mg/day orange juice leads to a statistically significant ($p < 0.05$) increase in length than vitamin C by itself. At 2.0 mg/day there is no significant difference in the average length between the two supplement types. Having a significant separation at 0.5 and 1.0 mg/day, but little separation at 2.0 mg/day suggests that there is a saturation effect with the OJ.

This can also be seen by performing a linear regression using `geom_smooth(method = 'lm', formula = y ~ x)`. The shaded regions are 95% confidence intervals for the regression line.

Fig 6. Linear regression of Length as a Function of Dosage

Split by Supplement Type



As suggested by the earlier analysis, we only see overlap of the curves for the two supplement types at 2.0 mg/day.

With this noisy and small of a dataset it is difficult to say with any certainty if the relationship between length and dosage is linear or not, although it does certainly look part of a curve. More data points in each experimental group (to decrease the sd) and more dosage rates (to improve a curve fit) would be required to be able to describe more about the relationship between the variables.

Appendix

Code for Figure 4

```
ggplot(data = ToothGrowth) +
    geom_boxplot(mapping = aes(x = dose, y = len, color = supp)) +
    labs(x = "Dose (mg/day)",
         y = "Length",
         title = "Fig 4. The Effect of Vitamin C on Tooth Growth in Guinea Pigs",
         subtitle = "Odontoblast Length, by Supplement Type and Dosage") +
    scale_color_manual(name = 'Supplement Type',
                       values = c('red', 'blue'),
                       labels = c('Orange Juice', 'Vitamin C'))
```

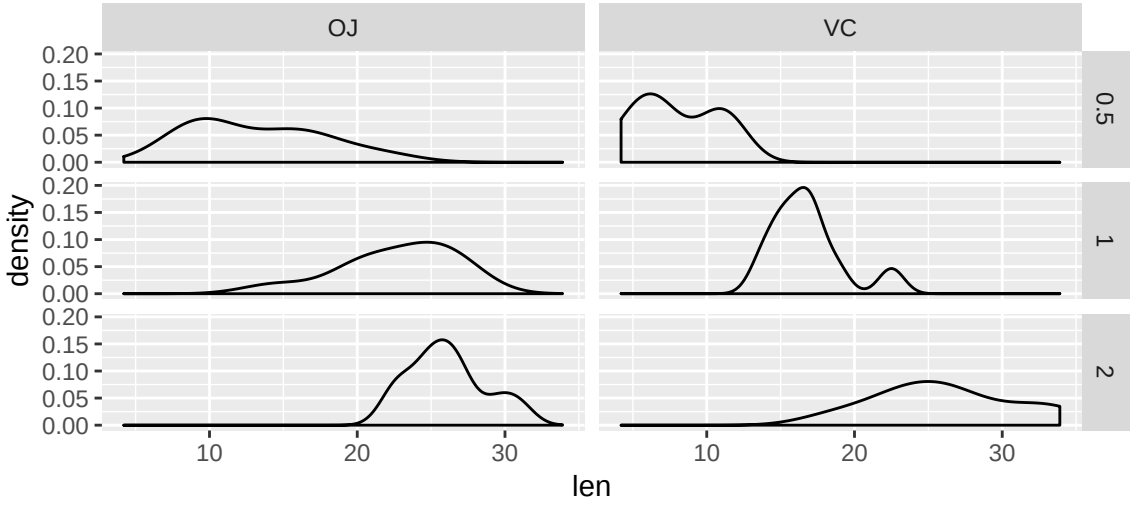
Code for Table 1

```
Tooth_summ <- ToothGrowth %>%
    group_by(dose, supp) %>%
    summarize("Average Length" = mean(len),
              "Length Std. Dev." = format(sd(len),digits = 3),
              "n" = n())
library(knitr)
kable(Tooth_summ,
      caption="Summary of ToothGrowth, grouped by dose and supplement",
      align = 'clccc',
      padding = 2)
```

Code for Figure 5

```
ggplot(data = ToothGrowth) +
    geom_density(mapping = aes(x = len)) +
    facet_grid(dose ~ supp) +
    labs(
        title = "Fig 5. Density plots of Length, split by Dose and Supplement Type"
    )
```

Fig 5. Density plots of Length, split by Dose and Supplement Type



###

Code for Table 2

```
Tooth <- data.frame(Dose = c(0, 0, 0), P.value = c(0, 0, 0))
j = 0
for(i in c(0.5, 1, 2)) {
    j = j + 1
    foo <- t.test(ToothGrowth$len[ToothGrowth$dose == i & ToothGrowth$supp == "OJ"], ToothGrowth$len[To
    Tooth$Dose[j] <- i
    Tooth$P.value[j] <- round(foo, digits = 4)
}
kable(Tooth,
      padding = 2,
      caption = "Unpaired Two-sample T-Test Between Supplements at Each Dosage")
```

Code for Table 3

```
Tooth2 <- data.frame(Supplement = c("0.5 to 1.0", "1.0 to 2.0"),
                     OJ.P.value = c(0, 0),
                     VC.P.value = c(0, 0))

Tooth2[1, 2] <- t.test(ToothGrowth$len[ToothGrowth$dose == 0.5 &
                                       ToothGrowth$supp == "OJ"],
                       ToothGrowth$len[ToothGrowth$dose == 1 &
                                       ToothGrowth$supp == "OJ"])$p.value
Tooth2[1, 3] <- t.test(ToothGrowth$len[ToothGrowth$dose == 0.5 &
                                       ToothGrowth$supp == "VC"],
                       ToothGrowth$len[ToothGrowth$dose == 1 &
                                       ToothGrowth$supp == "VC"])$p.value
Tooth2[2, 2] <- t.test(ToothGrowth$len[ToothGrowth$dose == 1 &
                                       ToothGrowth$supp == "OJ"],
                       ToothGrowth$len[ToothGrowth$dose == 2 &
                                       ToothGrowth$supp == "OJ"])$p.value
Tooth2[2, 3] <- t.test(ToothGrowth$len[ToothGrowth$dose == 1 &
                                       ToothGrowth$supp == "VC"],
                       ToothGrowth$len[ToothGrowth$dose == 2 &
                                       ToothGrowth$supp == "VC"])$p.value
kable(format(Tooth2, scientific = FALSE, digits = 1),
      padding = 2,
      caption = "Unpaired Two-sample T-Test Between Dosages for Each Supplement")
```

Code for Figure 6

```
ggplot(data = ToothGrowth,
       mapping = aes(x = as.numeric(as.character(dose)),
                     y = len,
                     color = supp)) +
    geom_point() +
    geom_smooth(method = 'lm', formula = y ~ x) +
    labs(x = "Dose (mg/day)",
         y = "Average Length",
         title = "Fig 6. Linear regression of Length as a Function of Dosage",
         subtitle = "Split by Supplement Type") +
    scale_color_manual(name = 'Supplement Type',
```

```
                    values = c('red', 'blue'),
                    labels = c('Orange Juice', 'Vitamin C'))
```

Fig 6. Linear regression of Length as a Function of Dosage

Split by Supplement Type