JHU Statistical Inference Basic Inferential Data Analysis

Overview

The ToothGrowth dataset represents a study of the effect of Vitamin C on tooth growth in 60 guinea pigs

The length of odontoblasts (cells responsible for tooth growth) is the response variable. Each animal was given one of three vitamin C dose levels (0.5, 1, and 2 mg/day) via one of two delivery methods: orange juice or ascorbic acid (a form of vitamin C and coded as VC), making 10 observations for each dosage/supplement combination.

The report finds that dosage levels have a significant impact on odontoblast length. Looking at all dosage levels, the supplement type could not be determined to be significant, however, breaking the data down to dosage levels, it finds that there is significant impact at dosage levels of 0.5 and 1.0 mg/day.

Where not included inline, all code for the following report can be found in the Appendix.

Data

The ToothGrowth dataset can be accessed via the `data` function in R and consists of 60 observations on 3 variables.

- len numeric Tooth length
- supp factor Supplement type (VC or OJ).
- dose numeric Dose in milligrams.

```
data(ToothGrowth); str(ToothGrowth)
```

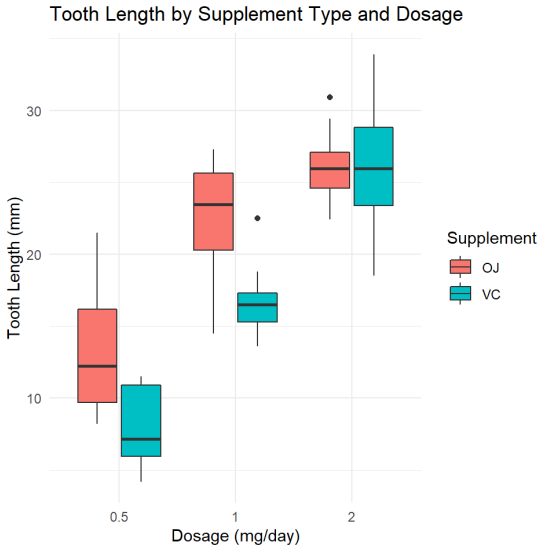
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

For this study, we will convert the dose variable to a discrete factor.

```
ToothGrowth <- ToothGrowth %>% mutate(dose = as.factor(dose))
```

Exploratory Data Analysis

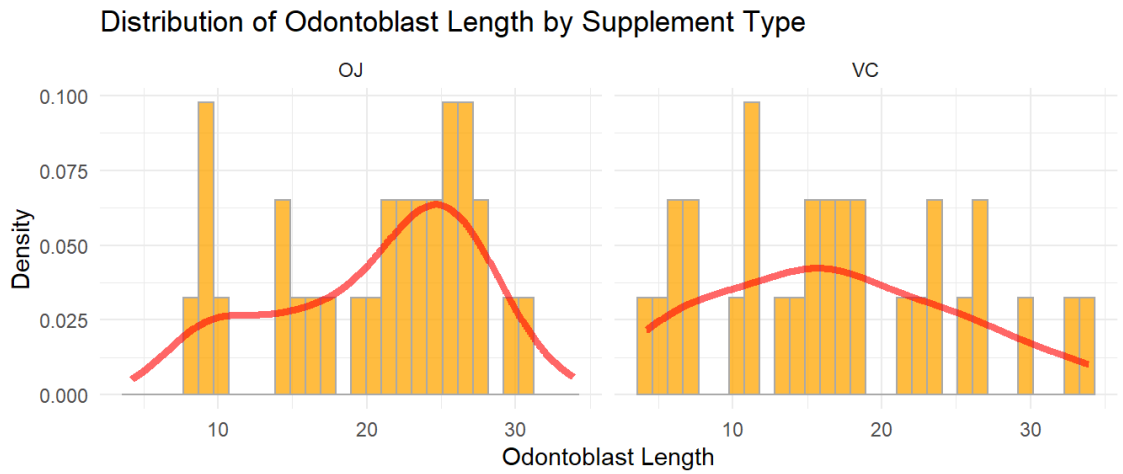
We start with a look at how the odontoblast length is distributed with respect to dosage levels and supplement types.

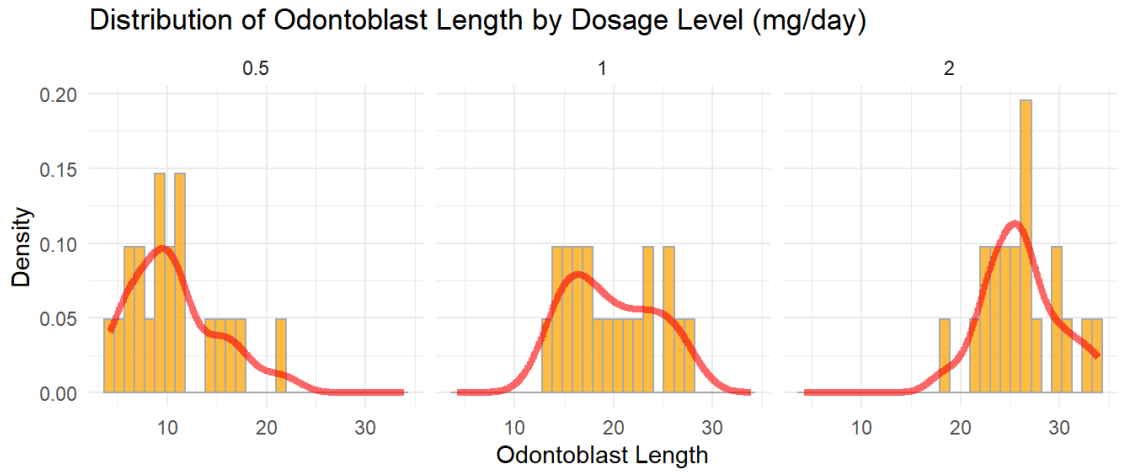


supp	dose	count	min	q25	median	q75	max	mean	sd	skew
OJ	0.5	10	8.2	9.700	12.25	16.175	21.5	13.23	4.46	0.51
OJ	1	10	14.5	20.300	23.45	25.650	27.3	22.70	3.91	-0.80
OJ	2	10	22.4	24.575	25.95	27.075	30.9	26.06	2.66	0.43
VC	0.5	10	4.2	5.950	7.15	10.900	11.5	7.98	2.75	0.16
VC	1	10	13.6	15.275	16.50	17.300	22.5	16.77	2.52	1.08
VC	2	10	18.5	23.375	25.95	28.800	33.9	26.14	4.80	0.19

- standard deviations are high, this is expected with such small sample sizes, and vary across all categories.
- orange juice appears to be more effective for dosages of 0.5 & 1mg/day but little difference at 2mg/day other than having a lower spread of results.
- both supplement types indicate a strong relationship with dosage quantity.
- a small amount of skew is visible amongst certain categories, however this should be within the tolerance of the t-test

Next, we will look at the distribution of odontoblast length for each variable to check for normality:





While none of the results show a classic Gaussian curve, distributions are close enough to normal for a T-analysis.

Analysis

Assumptions

- Each guinea pig received only one treatment type at one dosage level. Therefore, each observation is independent of the others and is not considered paired.
- Variances for each dosage level are 20.25, 19.5 & 14.24 and as such are considered unequal for the T test.
- Variances for each supplement type are 43.63 & 68.33 and as such are considered unequal for the T test.
- For the purposes of the analysis, it's assumed that apart from supplement and dosage of Vitamin C, all other factors affecting odontoblast growth are held equal across all subjects.

T-Test on Dosage

We will test two hypotheses: that average odontoblast length is equal for dosages 0.5 & 1.0, and similarly for dosages 1.0 & 2.0.

$H_{01} : \mu_{0.5} = \mu_{1.0}, H_{A1} : \mu_{0.5} \neq \mu_{1.0}$

$H_{02} : \mu_{0.5} = \mu_{1.0}, H_{A2} : \mu_{0.5} \neq \mu_{1.0}$

Summary

	P-value	CI Lower	CI Upper
0.5:1.0	1.00e-07	-11.983781	-6.276219
1.0:2.0	1.91e-05	-8.996481	-3.733519

Comparing odontoblast lengths for dosages of 0.5 & 1.0mg/day, and also for 1.0 & 2.0mg/day, there is significant difference on both intervals to reject H_{01} & H_{02}. Both tests report very low P-values and 95% confidence intervals that do not include zero.

T-Test on Dosage

We will test four hypotheses: that average odontoblast length is equal for vitamin C and ascorbic acid regardless of dosage level.
We will also perform the same test for dosages 0.5, 1.0, and 2.0.

$H_{01} : \mu_{OJ} = \mu_{VC}$, $H_{A1} : \mu_{OJ} \neq \mu_{VC}$

$H_{02} : \mu_{OJ0.5} = \mu_{VC0.5}$, $H_{A2} : \mu_{OJ0.5} \neq \mu_{VC0.5}$

$H_{03} : \mu_{OJ1.0} = \mu_{VC1.0}$, $H_{A3} : \mu_{OJ1.0} \neq \mu_{VC1.0}$

$H_{04} : \mu_{OJ2.0} = \mu_{VC2.0}$, $H_{A4} : \mu_{OJ2.0} \neq \mu_{VC2.0}$

Summary

	P-value	CI Lower	CI Upper
All dosages	0.0606345	-0.1710156	7.571016
Dosage 0.5	0.0063586	1.7190573	8.780943
Dosage 1.0	0.0010384	2.8021482	9.057852
Dosage 2.0	0.9638516	-3.7980705	3.638070

Across all dosages, although there is probable cause to believe that supplement type has an affect on odontoblast length, at the 95% confidence level, we are unable to reject the null hypothesis with a p-value of 0.06 and a confidence interval that includes zero.

Drilling down into the dosage levels however, we see that dosage levels at 0.5 and 1.0 mg/day do show a significant difference and only the higher dosage of 2.0 shows insignificant difference between supplements.

Therefore, we reject H_{02} & H_{03} while failing to reject H_{04}.

Conclusion

Based on this analysis, there is statistically significant difference for test groups receiving 0.5, 1.0 and 2.0 mg/day of supplement to say that in similar samples, by increasing dosage from 0.5 to 1.0 mg/day and from 1.0 to 2.0mg/day we would expect to see an increase in odontoblast length in at least 95% of test subjects.

While there was strong suggestion of a similar relationship between supplement types across all subjects, with 94% seeing an increase for Orange Juice, we fail to reject the null hypothesis that there is no effect.

Examining effects of different supplement types at each dosage level, a clearer picture emerges, where we see that orange juice is significantly different at 0.5 and 1.0 mg/day (99.4% and 99.0% respectively) but almost no difference at 2.0 mg/day (only 4%). We would therefore reject the null hypothesis at dosage levels of 0.5 and 1.0 mg/day and fail to reject at 2.0 mg/day. In other words, orange juice is considered more effective at lower dosage levels, but at 2.0 mg/day, there is no significant difference.

Further Studies

There are a lot of unknowns about the test subjects:

- are they genetically (and otherwise) similar?
- have they been exposed to similar environmental factors prior to the test?
- no control group was present in the study - what is the variance of untreated guinea pigs?

Additionally, with only 10 subjects per vector, it's very difficult to make any generalizations about the population.

Future studies would include a control group and greater sample size or many more sample sets.

Appendix

Included Libraries

This report uses the following libraries: `dplyr`, `tidyr`, `ggplot2`, `kableExtra`, `moments`

Code Blocks

Exploratory Data Analysis

Code for Exploratory Data Analysis & Data Summary Table

```
# boxplot of len, dose on x axis, split by supp
p <- ToothGrowth %>%
    ggplot(aes(y=len, x=as.factor(dose), fill=supp)) +
    geom_boxplot() +
    ggtitle("Tooth Length by Supplement Type and Dosage") +
    xlab("Dosage (mg/day)") +
    ylab("Tooth Length (mm)") +
    guides(fill=guide_legend(title="Supplement")) +
    theme_minimal()

# create kable of stat summaries
tg_summary <- ToothGrowth %>% group_by(supp, dose) %>%
    summarise(
        count=n(),
        min=min(len),
        q25=quantile(len, 0.25),
        median=median(len),
        q75=quantile(len, 0.75),
        max=max(len),
        mean=round(mean(len),2),
        sd=round(sd(len),2),
        skew=round(skewness(len),2)) %>%
    kbl() %>%
    kable_styling(bootstrap_options = c("condensed"))
```

Code for histograms

```
# histogram of len by supp with overlaid density curve
gsupp <- ToothGrowth %>% ggplot(aes(x=len)) +
    geom_histogram(alpha = 0.75, bins = 30, aes(y = stat(density)),
                   position="identity", fill="orange", col="darkgrey") +
    geom_line(aes(y = ..density..), colour = 'red', stat = 'density',
              size = 1.5, alpha = 0.6) + facet_wrap(~supp) +
    ggtitle ("Distribution of Odontoblast Length by Supplement Type") +
    xlab("Odontoblast Length") + ylab("Density") + theme_minimal()

# histogram of len by dose with overlaid density curve
gdose <- ToothGrowth %>% ggplot(aes(x=len)) +
    geom_histogram(alpha = 0.75, bins = 30, aes(y = stat(density)),
                   position="identity", fill="orange", col="darkgrey") +
    geom_line(aes(y = ..density..), colour = 'red', stat = 'density',
              size = 1.5, alpha = 0.6) + facet_wrap(~dose) +
    ggtitle ("Distribution of Odontoblast Length by Dosage Level (mg/day)") +
    xlab("Odontoblast Length") + ylab("Density") + theme_minimal()
```

Analysis

Variations of `len` for `dose` and `supp`

```
# get variation of len by dose and supp to test if vars are equal or not
varDose <- ToothGrowth %>% group_by(dose) %>% summarise(round(var(len),2))
varSupp <- ToothGrowth %>% group_by(supp) %>% summarise(round(var(len),2))
```

T-Test for `dose`

```
# run 2 t-tests for each dose interval (0.5:1, 1:2)
td1 <- t.test(len ~ dose, paired = FALSE, var.equal = FALSE,
              data = subset(ToothGrowth, dose %in% c(0.5, 1)))
td2 <- t.test(len ~ dose, paired = FALSE, var.equal = FALSE,
              data = subset(ToothGrowth, dose %in% c(1, 2)))

# create dataframe of useful stats
df_dose <- data.frame(td1$p.value, td1$conf.int[1], td1$conf.int[2])
df_dose <- rbind(df_dose, c(td2$p.value, td2$conf.int[1], td2$conf.int[2]))

rownames(df_dose) = c("0.5:1.0","1.0:2.0")

# create kable from summary
df_dose %>%
    kbl(col.names = c("P-value", "CI Lower", "CI Upper")) %>%
    kable_styling(bootstrap_options = c("condensed"),
                  full_width = F, position = "float_left")
```

T-Test for `supp`

```
# run 4 t-tests for supp: all dosages, dosage=0.5, dosage=1, dosage=2
tsupp <- t.test(len ~ supp, paired = FALSE, var.equal = FALSE,
                data = ToothGrowth)
tsuppd1 <- t.test(len ~ supp, paired = FALSE, var.equal = FALSE,
                  data = subset(ToothGrowth, dose == 0.5))
tsuppd2 <- t.test(len ~ supp, paired = FALSE, var.equal = FALSE,
                  data = subset(ToothGrowth, dose == 1))
tsuppd3 <- t.test(len ~ supp, paired = FALSE, var.equal = FALSE,
                  data = subset(ToothGrowth, dose == 2))

# create dataframe of useful stats
df_supp <- data.frame(tsupp$p.value, tsupp$conf.int[1], tsupp$conf.int[2])
df_supp <- rbind(df_supp,
                 c(tsuppd1$p.value, tsuppd1$conf.int[1], tsuppd1$conf.int[2]))
df_supp <- rbind(df_supp,
                 c(tsuppd2$p.value, tsuppd2$conf.int[1], tsuppd2$conf.int[2]))
df_supp <- rbind(df_supp,
                 c(tsuppd3$p.value, tsuppd3$conf.int[1], tsuppd3$conf.int[2]))

rownames(df_supp) = c("All dosages", "Dosage 0.5", "Dosage 1.0", "Dosage 2.0")

# create kable from summary
df_supp %>%
    kbl(col.names = c("P-value", "CI Lower", "CI Upper")) %>%
    kable_styling(bootstrap_options = c("condensed"),
                  full_width = F, position = "float_left")
```