Inferential analysis on Tooth Growth response for Guinea Pigs

Syed Abdullah Hasan

6/20/2021

Overview

This report is prepared as the final project submission for the Statistical Inference course offered by Johns Hopkins University on Coursera. This second part of the report analyses Tooth Growth data from the R package to infer relationships between tooth growth and supplement dosage, supported by hypothesis testing and conclusions. The report concludes that dosage significantly affects the tooth growth response rate in Guinea Pigs, but the choice of supplement is does not have a statistically significant impact beyond a dosage of 1.0 mg/day.

Part 2 - Basic Inferential Data Analysis

This section of the report will provide a basic inferential data analysis on the Tooth Growth data in the R package. The data is first loaded into an R object for processing.

```
data("ToothGrowth")
ToothGrowth$dose <- as.factor(as.character(ToothGrowth$dose))
```

Data Summary

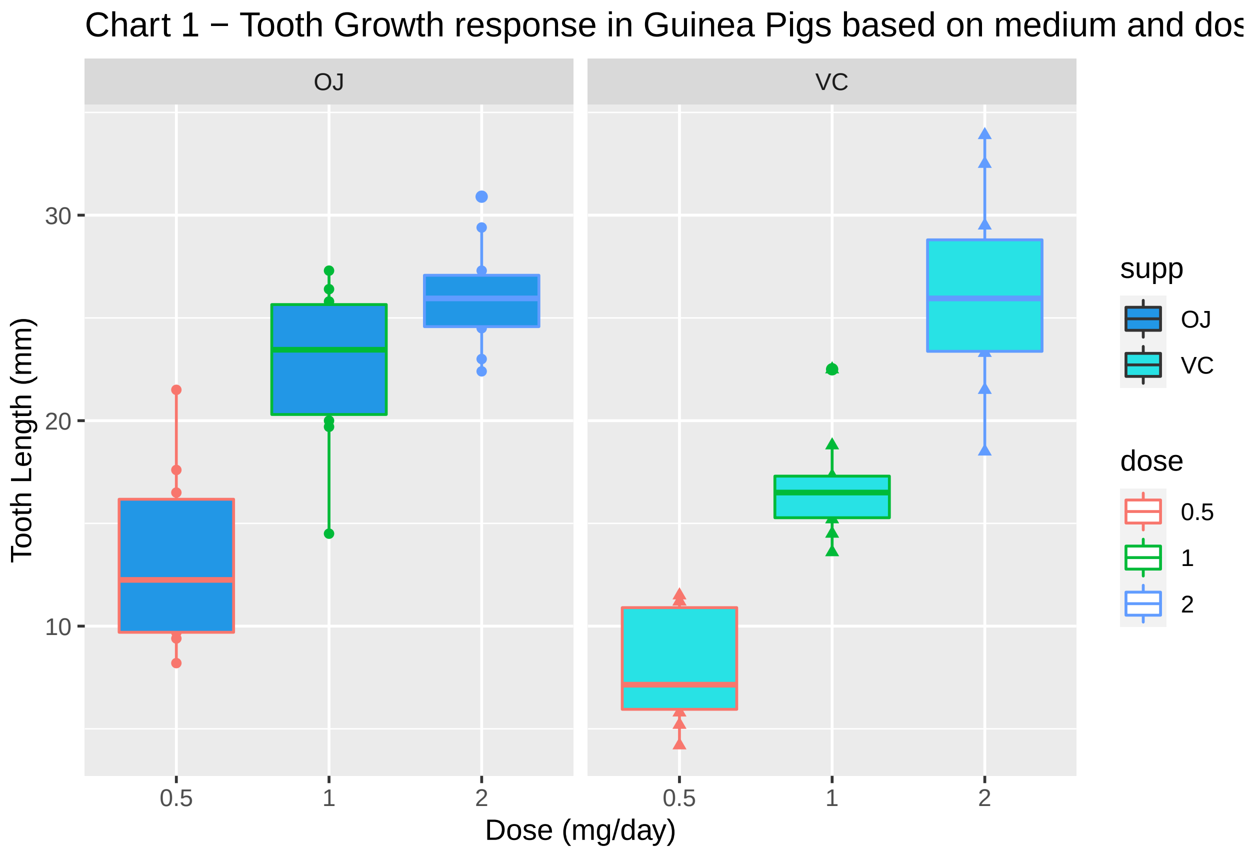
This data set covers the results of testing on Guinea Pigs to evaluate the effect of Vitamin C on tooth growth. The data set includes three variables:

- "len" denoting the response, which is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs;
- "supp" denoting the supplement administered (in this case Vitamin C) ; and,
- "dose" denoting the level of Vitamin C administered. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice (coded as OJ) or ascorbic acid (a form of vitamin C and coded as VC).

```
g1 <- ggplot(ToothGrowth, aes(x = dose, y = len,
                              color = dose,
                              group = dose,
                              shape = supp)) +
        facet_wrap(~supp) +
        geom_point () +
        geom_boxplot(aes(fill=supp)) +
        scale_fill_manual (values = c(4,5)) +
        labs (title = "Chart 1 - Tooth Growth response in Guinea Pigs based on medium and dosage",
              x = "Dose (mg/day)",
```

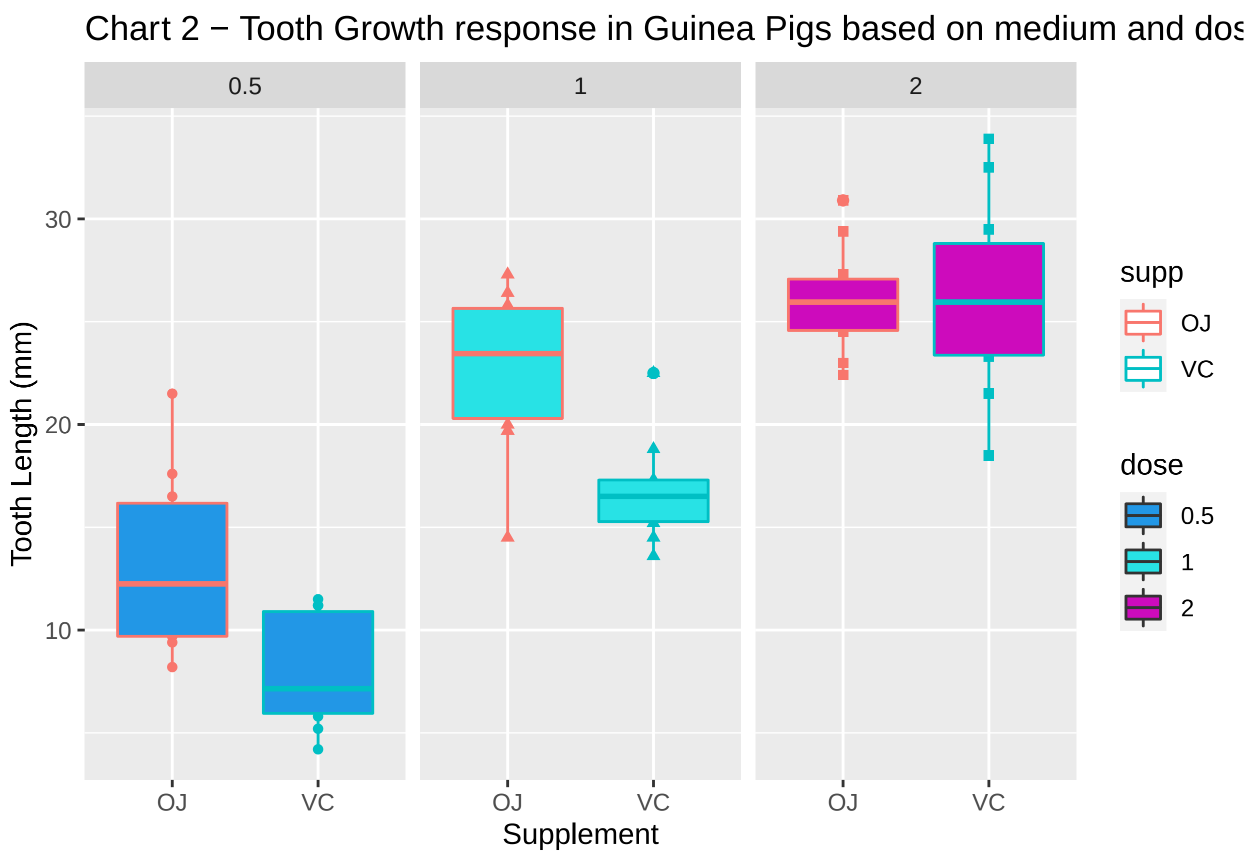
```
            y = "Tooth Length (mm)")

g1
```




```
g2 <- ggplot(ToothGrowth, aes(x = supp, y = len,
                              color = supp,
                              group = supp,
                              shape = dose)) +
        facet_wrap(~dose) +
        geom_point () +
        geom_boxplot(aes(fill=dose)) +
        scale_fill_manual (values = c(4,5,6)) +
        labs (title = "Chart 2 - Tooth Growth response in Guinea Pigs based on medium and dosage",
            x = "Supplement",
            y = "Tooth Length (mm)")

g2
```




```
df_summary <-  ToothGrowth %>% summarise("Supplement" = "All",
                                         "Dose" = "All",
                                         "Mean" = mean(len),
                                         "Std." = sd(len),
                                         "Min." = min(len),
                                         "Q1" = quantile(len,0.25),
                                         "Median" = median(len),
                                         "Q3" = quantile(len,0.75),
                                         "Max." = max(len))

df_summary <-  rbind(df_summary, ToothGrowth %>% group_by ("Supplement" = supp) %>%
                                summarise("Dose" = "All",
                                "Mean" = mean(len),
                                "Std." = sd(len),
                                "Min." = min(len),
                                "Q1" = quantile(len,0.25),
                                "Median" = median(len),
                                "Q3" = quantile(len,0.75),
                                "Max." = max(len))
                     )

df_summary <-  rbind(df_summary,
                     ToothGrowth %>%
                     group_by ("Supplement" = "All",
                               "Dose" = as.character(format(dose,nsmall=1))) %>%
                               summarise("Mean" = mean(len),
```

```
                                  "Std." = sd(len),
                                  "Min." = min(len),
                                  "Q1" = quantile(len,0.25),
                                  "Median" = median(len),
                                  "Q3" = quantile(len,0.75),
                                  "Max." = max(len))
                       )


df_summary <-  rbind(df_summary,
                     ToothGrowth %>%
                     group_by ("Supplement" = supp,
                               "Dose" = as.character(format(dose,nsmall=1))) %>%
                               summarise("Mean" = mean(len),
                                         "Std." = sd(len),
                                         "Min." = min(len),
                                         "Q1" = quantile(len,0.25),
                                         "Median" = median(len),
                                         "Q3" = quantile(len,0.75),
                                         "Max." = max(len))
                     )


kable (df_summary,
       caption = "Table 1 - Summary statistics for Tooth Growth data set",
       digits = 2,
       format.args = list(big.mark=",",
                       scientific = FALSE,
                       nsmall = 2
                               )
       )
```

Table 1: Table 1 - Summary statistics for Tooth Growth data set

Supplement	Dose	Mean	Std.	Min.	Q1	Median	Q3	Max.
All	All	18.81	7.65	4.20	13.07	19.25	25.27	33.90
OJ	All	20.66	6.61	8.20	15.52	22.70	25.73	30.90
VC	All	16.96	8.27	4.20	11.20	16.50	23.10	33.90
All	0.5	10.61	4.50	4.20	7.22	9.85	12.25	21.50
All	1	19.74	4.42	13.60	16.25	19.25	23.38	27.30
All	2	26.10	3.77	18.50	23.53	25.95	27.83	33.90
OJ	0.5	13.23	4.46	8.20	9.70	12.25	16.18	21.50
OJ	1	22.70	3.91	14.50	20.30	23.45	25.65	27.30
OJ	2	26.06	2.66	22.40	24.58	25.95	27.08	30.90
VC	0.5	7.98	2.75	4.20	5.95	7.15	10.90	11.50
VC	1	16.77	2.52	13.60	15.27	16.50	17.30	22.50
VC	2	26.14	4.80	18.50	23.38	25.95	28.80	33.90

As seen in Chart 4 and Table 1, the response rate for tooth growth in Guinea Pigs appears to be greatest in the case of Vitamin C when administered 2.0 mg/day. However, the variability in this case is also very high. For lower doses of Vitamin C, the response rate appears to be weaker than when Orange Juice is administered. However, variability in growth rates when Orange Juice is administered appears to be higher when 0.5 mg/day or 1.0 mg/day is administered.

Hypothesis Formulation and Testing

Based on these initial observations, the key questions to be evaluated across the data set are as follows:

- Case 1: Does Orange Juice lead to higher tooth growth as compared with Vitamin C?
- Case 2: Does higher dosage lead to higher tooth growth?
- Case 3: Does choice of supplement contribute to higher tooth growth for a given dosage level?

The following hypothesis may be developed for testing the above cases:

Testing effect of supplement on growth response Case 1. Administering Orange Juice leads to a higher response rate versus Vitamin C:
- H_0: μ_{OJ} - $\mu_{VC} = 0$
- H_1: μ_{OJ} - $\mu_{VC} <> 0$

Testing effect of dosage on growth response Case 2a. Administering 1.0 mg/day dosage leads to a higher response rate versus 0.5 mg/day:
- H_0: $\mu_{D=0.5}$ - $\mu_{D=1.0} = 0$
- H_1: $\mu_{D=0.5}$ - $\mu_{D=1.0} <> 0$

Case 2b. Administering 2.0 mg/day dosage leads to a higher response rate versus 0.5 mg/day:
- H_0: $\mu_{D=0.5}$ - $\mu_{D=2.0} = 0$
- H_1: $\mu_{D=0.5}$ - $\mu_{D=2.0} <> 0$

Case 2c. Administering 2.0 mg/day dosage leads to a higher response rate versus 1.0 mg/day:
- H_0: $\mu_{D=0.5}$ - $\mu_{D=2.0} = 0$
- H_1: $\mu_{D=0.5}$ - $\mu_{D=2.0} <> 0$

Testing supplement as a factor for given dose levels Case 3a. Administering Orange Juice leads to a higher response rate for doses of 0.5 mg/day:
- H_0: $\mu_{OJ} \mid$ ~D = 0.5~ - $\mu_{VC} \mid$ ~D = 0.5~ = 0
- H_1: $\mu_{OJ} \mid$ ~D = 0.5~ - $\mu_{VC} \mid$ ~D = 0.5~ <> 0

Case 3b. Administering Orange Juice leads to a higher response rate for doses of 1.0 mg/day:
- H_0: $\mu_{OJ} \mid$ ~D = 1.0~ - $\mu_{VC} \mid$ ~D = 1.0~ = 0
- H_1: $\mu_{OJ} \mid$ ~D = 1.0~ - $\mu_{VC} \mid$ ~D = 1.0~ <> 0

Case 3c. Administering Orange Juice leads to a higher response rate for doses of 2.0 mg/day:
- H_0: $\mu_{OJ} \mid$ ~D = 2.0~ - $\mu_{VC} \mid$ ~D = 2.0~ = 0
- H_1: $\mu_{OJ} \mid$ ~D = 2.0~ - $\mu_{VC} \mid$ ~D = 2.0~ <> 0

The data is first subset to evaluate each of the cases mentioned above.

```
c2a <- filter(ToothGrowth, dose %in% c(0.5, 1.0))
c2b <- filter(ToothGrowth, dose %in% c(0.5, 2.0))
c2c <- filter(ToothGrowth, dose %in% c(1.0, 2.0))
c3a <- filter(ToothGrowth, dose == 0.5)
c3b <- filter(ToothGrowth, dose == 1.0)
c3c <- filter(ToothGrowth, dose == 2.0)
```

A t-test may be used to evaluate the null hypothesis, since the sample size is small in all cases. The data is not paired since a single observation is recorded for each of 60 subjects. However, variances between the factors need to be tested in order to confirm the parameters for the t-test.

```
v1 <- var.test (len ~ supp, ToothGrowth)
v2a <- var.test (len ~ dose, c2a)
v2b <- var.test (len ~ dose, c2b)
v2c <- var.test (len ~ dose, c2c)
v3a <- var.test (len ~ supp, c3a)
v3b <- var.test (len ~ supp, c3b)
v3c <- var.test (len ~ supp, c3c)

table <- map_df(list(v1, v2a, v2b, v2c, v3a, v3b, v3c), tidy)[,c(2,4,5,8,9)]
```

```
## Multiple parameters; naming those columns num.df, den.df
## Multiple parameters; naming those columns num.df, den.df
## Multiple parameters; naming those columns num.df, den.df
## Multiple parameters; naming those columns num.df, den.df
## Multiple parameters; naming those columns num.df, den.df
## Multiple parameters; naming those columns num.df, den.df
## Multiple parameters; naming those columns num.df, den.df
```

```
names(table) <- c("Parameter (df)", "Statistic", "p-value",
                  "Method", "Alternative")
row.names(table) <- c("Case 1", "Case 2a", "Case 2b", "Case 2c", "Case 3a",
                      "Case 3b", "Case 3c")
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
kable (table,
       digits = 2,
       format.args = list(big.mark=",",
                          scientific = FALSE,
                          nsmall = 2),
       caption = "Table 2 - Summary of F-test to compare sample variances for hypothesis testing"
        )
```

Table 2: Table 2 - Summary of F-test to compare sample variances for hypothesis testing

	Parameter (df)	Statistic	p-value	Method	Alternative
Case 1	29.00	0.64	0.23	F test to compare two variances	two.sided
Case 2a	19.00	1.04	0.94	F test to compare two variances	two.sided
Case 2b	19.00	1.42	0.45	F test to compare two variances	two.sided
Case 2c	19.00	1.37	0.50	F test to compare two variances	two.sided
Case 3a	9.00	2.64	0.16	F test to compare two variances	two.sided
Case 3b	9.00	2.42	0.20	F test to compare two variances	two.sided
Case 3c	9.00	0.31	0.09	F test to compare two variances	two.sided

Based on the p-values in Table 2 above, we can accept the null hypotheses that the true ratio of variances is equal to 1 in all cases. The parameters for t-testing will therefore be set to incorporate equal variances.

```r
t1 <- t.test (len ~ supp, data=ToothGrowth,
              conf.level = 0.95, alternative = "two.sided", var.equal = T)
t2a <- t.test (len ~ dose, data=c2a, conf.level = 0.95,
               alternative = "two.sided", var.equal = T)
t2b <- t.test (len ~ dose, data=c2b, conf.level = 0.95,
               alternative = "two.sided", var.equal = T)
t2c <- t.test (len ~ dose, data=c2c, conf.level = 0.95,
               alternative = "two.sided", var.equal = T)
t3a <- t.test (len ~ supp, data=c3a, conf.level = 0.95,
               alternative = "two.sided", var.equal = T)
t3b <- t.test (len ~ supp, data=c3b, conf.level = 0.95,
               alternative = "two.sided", var.equal = T)
t3c <- t.test (len ~ supp, data=c3c, conf.level = 0.95,
               alternative = "two.sided", var.equal = T)

table <- map_df(list(t1, t2a, t2b, t2c, t3a, t3b, t3c), tidy)[,4:10]
names(table) <- c("Statistic", "p-value", "Parameter (df)",
                  "C.I. - Low", "C.I. - High", "Method", "Alternative")
row.names(table) <- c("Case 1", "Case 2a", "Case 2b", "Case 2c", "Case 3a",
                      "Case 3b", "Case 3c")
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```r
kable (table,
       digits = 2,
       format.args = list(big.mark=",",
                          scientific = FALSE,
                          nsmall = 2),
       caption = "Table 3 - Summary of t-test to evaluate hypothesis for each case"
        )
```

Table 3: Table 3 - Summary of t-test to evaluate hypothesis for each case

	Statistic	p-value	Parameter (df)	C.I. - Low	C.I. - High	Method	Alternative
Case 1	1.92	0.06	58.00	-0.17	7.57	Two Sample t-test	two.sided
Case 2a	-6.48	0.00	38.00	-11.98	-6.28	Two Sample t-test	two.sided
Case 2b	-11.80	0.00	38.00	-18.15	-12.84	Two Sample t-test	two.sided
Case 2c	-4.90	0.00	38.00	-8.99	-3.74	Two Sample t-test	two.sided
Case 3a	3.17	0.01	18.00	1.77	8.73	Two Sample t-test	two.sided
Case 3b	4.03	0.00	18.00	2.84	9.02	Two Sample t-test	two.sided
Case 3c	-0.05	0.96	18.00	-3.72	3.56	Two Sample t-test	two.sided

Based on the summary statistics and p-values in Table 3 above, we can accept the null hypothesis for case 1 and 3c. In all other cases, we can reject the null hypothesis that the true difference in means between two variables is zero.

Conclusions and supporting assumptions

The report can finally conclude the following statistically significant relationships:

- Case 1: Administering Orange juice does not significantly boost tooth growth as compared with Vitamin C.
- Case 2a-c: A higher dosage of supplement significantly boosts tooth growth, regardless of supplement used.
- Case 3a-c: Up to dosage of 1.0 mg/day, administering Orange Juice significantly leads to a higher response rate for tooth growth - however, this relationship is not statistically significant for a dosage of 2.0mg/day.

In order to arrive at the above conclusions, the following assumptions for two-sample t-tests are made:

- The response rate data is continuous and normally distributed.
- The variances of the two populations (those administered Vitamin C and those administered Orange Juice) are equal.
- All subjects included in the study are randomly sampled and independent - therefore, there is no prior underlying relationship between the response rate and the subjects.