Statistical Inference Course Project: Basic Inferential Data Analysis

Tushar Kataria

Contents

Overview

The second part of the course project involves analysing the `ToothGrowth` data in the R `datasets` package which describes the effect of vitamin C on tooth growth in guinea pigs, and performing basic exploratory data analyses on it. Following this, a series of hypothesis tests are utilised to assess the effectiveness of the vitamin C, focusing on how the vitamin was administered (the supplement type) and the dosage level used.

The `ToothGrowth` dataset is composed of 60 observations measuring three variables: `len` to record the total tooth length in millimetres (a numeric variable), `supp` for the supplement type as a factor with levels 'VC' and 'OJ' for ascorbic acid and orange juice, respectively; and `dose` to measure the dose in milligrams (as a numeric variable).

Overall, there were 60 guinea pigs that were separated into three groups, each of which received a specific dose of vitamin C (0.5, 1 or 2 mg). Within each group of 20 guinea pigs, one of two delivery methods were used - either ascorbic acid or orange juice - such that each subgroup contained 10 guinea pigs.

Exploratory Data Analysis

By loading and viewing the dataset (along with the necessary libraries), it is confirmed that there are 3 variables recording 60 observations which can be assumed to represent a population of 60 different guinea pigs divided into six subsets of 10 guinea pis each. These subgroups cover all the possible combinations of two delivery methods and three doses. Another assumption being made is that all of the guinea pigs were chosen randomly so that the population variance is the same for all subsets.

```r
data("ToothGrowth"); df <- ToothGrowth; str(df)
```

```
## '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 ...
```

```r
unique(df$dose)
```

```
## [1] 0.5 1.0 2.0
```

As there are only three levels of doses (0.5, 1 and 2 mg), the variable `dose` is converted to a factor type.

```r
df$dose <- factor(df$dose); str(df)
```

```
## '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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

After this data processing, `summary()` and `sd()` are used to provide a brief overview of the variables' statistics.
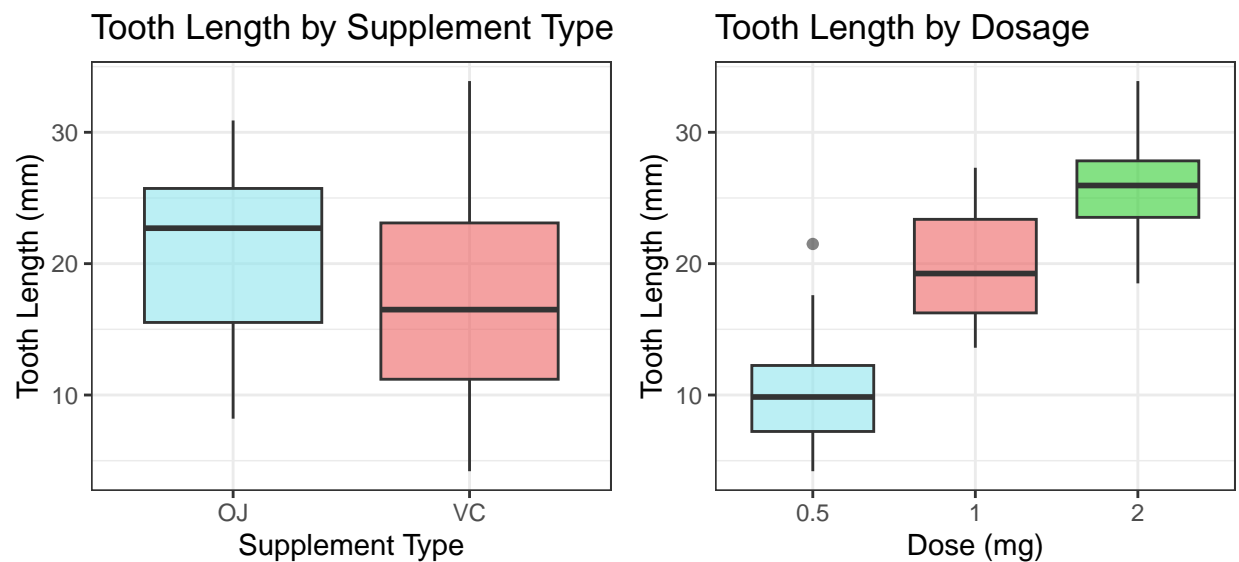
```
summary(df)
```

```
##       len          supp       dose
##  Min.   : 4.20   OJ:30    0.5:20
##  1st Qu.:13.07   VC:30    1  :20
##  Median :19.25            2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```
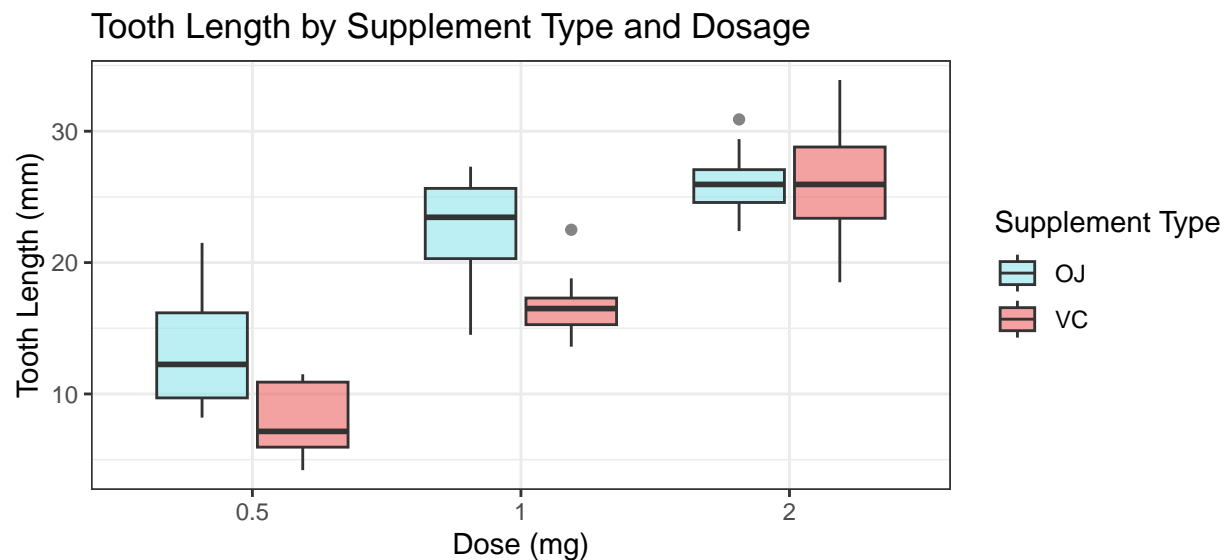
```
sd(df$len)
```

```
## [1] 7.649315
```

From this, the average guinea pig tooth length is 18.813 with a standard deviation of 7.649.

The left and right box plots above graphically present key statistics (the mean, the range, and the first and third quartiles) for the tooth length based on the delivery method and the dosage level, respectively.

Upon initial inspection, using orange juice results in longer tooth levels for 0.5 and 1 mg doses but is comparable to ascorbic acid for 2 mg. Furthermore, the tooth length generally increases as the dose increases. Additionally, the tooth length difference for 0.5 and 1 mg is larger than the length difference for 1 and 2 mg These are tested through t-tests; these t-tests are unpaired as the combinations of supplement types and doses are used on different guinea pigs.

Density Estimation by Supplement Type and Dosage
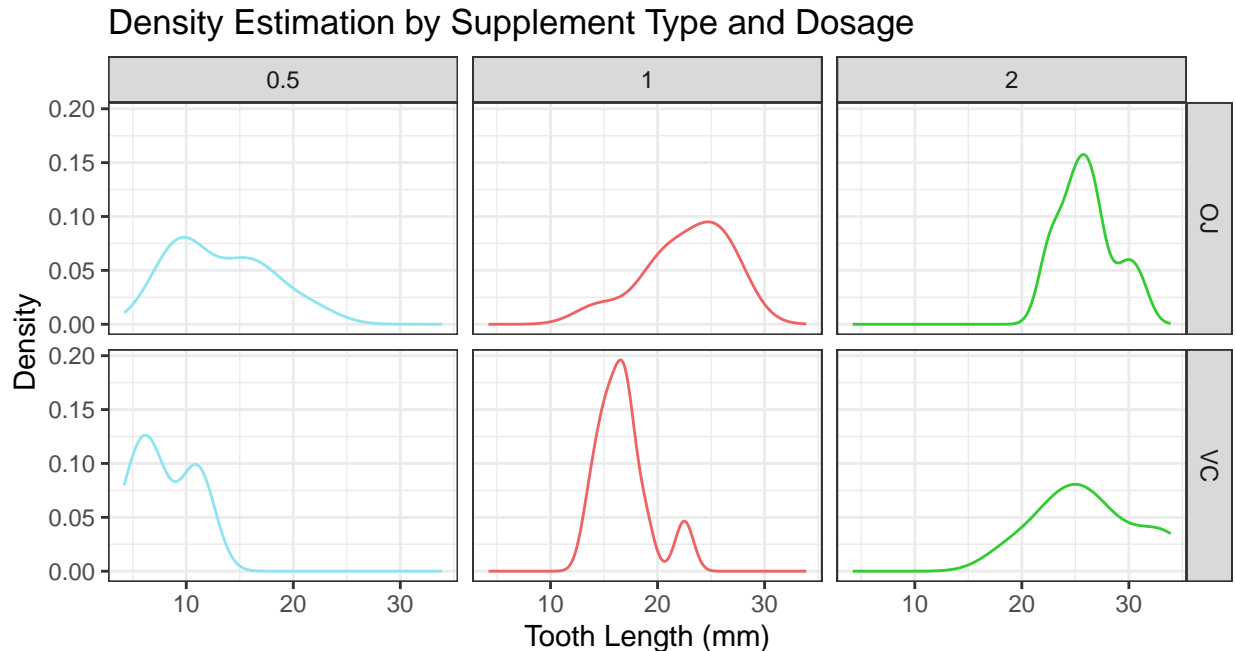
The trend of larger doses increasing the tooth length more than smaller doses is also suggested through the density estimation above. As the dose increases, the distribution becomes more right-skewed for both orange juice and ascorbic acid.

```
df %>%
  group_by(supp, dose) %>%
  summarize(mean = mean(len), .groups = "drop") %>%
  spread(supp, mean) %>%
  mutate(diff = abs(VC - OJ))
```

```
## # A tibble: 3 x 4
##   dose     OJ    VC   diff
##   <fct> <dbl> <dbl>  <dbl>
## 1 0.5    13.2  7.98 5.25
## 2 1      22.7 16.8  5.93
## 3 2      26.1 26.1  0.0800
```

The difference in means between VC and OJ for 2 mg is much smaller than the differences for other doses, so comparing the effectiveness between VC and OJ for 2 mg is more difficult. However, to formally test the effectiveness of the two, t-tests are used to calculate p-values and confidence intervals to perform hypothesis testing.

Hypothesis Testing

Assumptions

As the subgroups here have sample populations $n \leq 30$, the standard error estimate will generally not be accurate, so the t-distribution in constructing confidence intervals.

Also, several assumptions are made during this hypothesis testing:

- The samples are independent and identically distributed (i.i.d.) so the guinea pigs were randomly assigned to subgroups.
- The tooth length in all subgroups are normally distributed with an unknown mean and variance.
- The sample variances of tooth growth are different between different groups (combinations of supplement method and dosage level).

Also, the significance level is set at $\alpha = 0.05$ so any results with p-values below this are considered significant.

Variation of Tooth Length from Supplement Types

From the `Tooth Length by Supplement Type` box plot, it appears there is longer tooth length when using orange juice than ascorbic acid, so one-sided t-tests are used with the assumption that the two groups have unequal variances. Here, four tests are conducted depending on the dose level- across all doses, 0.5 mg, 1 mg and 2 mg. The null and alternative hypotheses for all of these are:

- $H_0 : \mu_{OJ} = \mu_{VC}$ - There is no significant difference in the mean tooth length between using OJ and VC.
- $H_a : \mu_{OJ} > \mu_{VC}$: There is a significant difference in the mean tooth length between using OJ and VC.

```
##                  H_0             H_a       dose           t        df      p.value
## 1 mu_OJ - mu_VC = 0 mu_OJ - mu_VC > 0 All doses   1.9152683 55.30943 0.060634508
## 2 mu_OJ - mu_VC = 0 mu_OJ - mu_VC > 0    0.5 mg   3.1697328 14.96875 0.006358607
## 3 mu_OJ - mu_VC = 0 mu_OJ - mu_VC > 0      1 mg   4.0327696 15.35767 0.001038376
## 4 mu_OJ - mu_VC = 0 mu_OJ - mu_VC > 0      2 mg  -0.0461361 14.03982 0.963851589
##    adjusted_p   conf.low conf.high est_mu_OJ est_mu_VC
## 1          NA -0.1710156  7.571016  20.66333  16.96333
## 2 0.009537910  1.7190573  8.780943  13.23000   7.98000
## 3 0.003115128  2.8021482  9.057852  22.70000  16.77000
## 4 0.963851589 -3.7980705  3.638070  26.06000  26.14000
```

Comparing OJ and VC across all dose levels, the p-value is $0.06 > 0.05$ and the 95% confidence interval $CI_{95\%} = [-0.17, 7.57]$ contains zero, so this fails to reject the null hypothesis. This is reflected in the `Tooth Growth by Supplement Type` plot above as the two box plots are close to each other. However, since the p-value is quite close to α, there is evidence in favor of H_0 is weak.

Due to multiple testing, arising from conducting individual t-tests for each dosage, the p-values are adjusted using the Benjamini-Hochberg method to control the False Discovery Rate (FDR). The adjusted p-values for 0.5 and 1 mg doses are smaller than α, so the null hypothesis is rejected for both: there is a significant difference in mean tooth length between using OJ and VC. The confidence intervals $CI_{95\%} = [1.72, 8.78]$ and $CI_{95\%} = [2.80, 9.06]$ for 0.5 mg and 1 mg, respectively, both show that this difference is positive and tooth length is greater for OJ as the supplement type. However, for 2 mg, the adjusted p-value of $0.96 \gg 0.05$ and 95% confidence interval $CI_{95\%} = [-3.80, 3.64]$ containing zero mean the data fails to reject the null hypothesis. This is highlighted in the `Tooth Length by Supplement Type and Dosage` plot where the box plot for OJ and VC are very similar for the 2 mg dose.

Variations of Tooth Length from Dose Levels

Again, multiple t-tests are carried out in order to cover all of the three pairs of doses. Since any difference in the mean lengths between doses is significant, positive or negative, two-tailed tests are conducted here as well. For these three combinations (0.5 mg vs. 1 mg, 0.5 mg vs. 2 mg, and 1 mg vs. 2 mg) the null and alternative hypotheses are as follows:

- $H_0 : \mu_{0.5,1,2mg} = \mu_{1,2,0.5mg}$: There is no significant difference in the mean tooth length between giving 0.5 mg and 1 mg.
- $H_a : \mu_{0.5,1,2mg} \neq \mu_{1,2,0.5mg}$: There is a significant difference in the mean tooth length between giving 0.5 mg and 1 mg.

```
##                 H_0               H_a          t       df      p.value
## 1 mu_1 - mu_0.5 = 0 mu_1 - mu_0.5 != 0   6.476648 37.98641 1.268301e-07
## 2 mu_2 - mu_0.5 = 0 mu_2 - mu_0.5 != 0  11.799046 36.88259 4.397525e-14
## 3   mu_2 - mu_1 = 0   mu_2 - mu_1 != 0   4.900484 37.10109 1.906430e-05
##      adjusted_p  conf.low conf.high est_mu_dose_x est_mu_dose_y
## 1 1.902451e-07  6.276219 11.983781        19.735        10.605
## 2 1.319257e-13 12.833833 18.156167        26.100        10.605
## 3 1.906430e-05  3.733519  8.996481        26.100        19.735
```

The original p-values here are also adjusted with the Benjamini-Hochberg method to control the FDR. In all three of the cases, the p-values and adjusted p-values are all much smaller than the set significance level of $\alpha = 0.05$ so the null hypothesis is rejected in every test. The data provides strong evidence that the dose level influences the mean tooth length in guinea pigs. This follows the initial supposition that the average tooth length increases with an increasing dose.

Conclusions

For this part of the project, the key assumptions made are that the samples (the 60 guinea pigs) are representative of the entire population and are independent and identically distributed (i.i.d.). This means the guinea pigs were randomly assigned to the supplement-dose subgroups. Another assumption is that the Central Limit Theorem (CLT) holds here and that the means of populations are normally distributed. Moreover, for the t-tests, it is assumed that the variances of different subgroups is different.

Overall, the results from the hypothesis tests and confidence intervals imply that an increase in dose implies and increase in average tooth length. The `Tooth Length by Supplement Type and Dosage` and `Density Estimation by Supplement Type and Dosage` plots show this being the case for both supplements. In terms of supplement type, there is not a statistically significant difference between orange juice ad ascorbic acid if the dose level is not also considered. However, in the case dosage is taken into account, orange juice is the better supplement type for 0.5 and 1 mg. But, for 2 mg, both methods are comparable and achieve similar results.

Appendix

Code for Figure 1

```
g1 <- ggplot(df, aes(x = supp, y = len, fill = supp)) +
  geom_boxplot(alpha = 0.6) +
  scale_fill_manual(values = c("cadetblue2", "indianred2")) +
  labs(title = "Tooth Length by Supplement Type",
       x = "Supplement Type",
       y = "Tooth Length (mm)") +
  theme_bw() +
  theme(legend.position = "none")

g2 <- ggplot(df, aes(x = dose, y = len, fill = dose)) +
  geom_boxplot(alpha = 0.6) +
  scale_fill_manual(values = c("cadetblue2", "indianred2", "limegreen")) +
  labs(title = "Tooth Length by Dosage",
       x = "Dose (mg)",
       y = "Tooth Length (mm)") +
  theme_bw() +
  theme(legend.position = "none")

grid.arrange(g1, g2, nrow = 1, ncol = 2)
```

Code for Figure 2

```
g <- ggplot(df, aes(x = dose, y = len, fill = supp)) +
  geom_boxplot(alpha = 0.6) +
  scale_fill_manual(values = c("cadetblue2", "indianred2")) +
  labs(title = "Tooth Length by Supplement Type and Dosage",
       x = "Dose (mg)",
       y = "Tooth Length (mm)",
       fill = "Supplement Type") +
  theme_bw() +
  theme(legend.position = "right")

g
```

Code for Figure 3

```
g <- ggplot(df, aes(x = len, color = dose)) +
  geom_density() +
  scale_color_manual(values = c("cadetblue2", "indianred2", "limegreen")) +
  facet_grid(supp ~ dose) +
  labs(title = "Density Estimation by Supplement Type and Dosage",
       x = "Tooth Length (mm)",
       y = "Density") +
  theme_bw() +
  theme(legend.position = "none")

g
```

Code for Supplement Type t-tests

```r
len_oj_all <- df$len[df$supp == "OJ"]
len_vc_all <- df$len[df$supp == "VC"]

test_all <- t.test(len_oj_all, len_vc_all,
                   paired = FALSE, var.equal = FALSE)

result_all <- test_all %>%
  tidy %>%
  mutate(H_0 = "mu_OJ - mu_VC = 0", H_a = "mu_OJ - mu_VC > 0",
         dose = "All doses",
         t = statistic, df = parameter,
         adjusted_p = NA,
         est_mu_OJ = estimate1, est_mu_VC = estimate2) %>%
  select(H_0, H_a, dose,
         t, df, p.value, adjusted_p,
         conf.low, conf.high,
         est_mu_OJ, est_mu_VC)

# print.data.frame(result)

#### Pairwise t-tests for different doses

dose_levels <- sort(unique(df$dose)) # Sorted in ascending order

pairwise_tests <- data.frame()

for (ds in dose_levels) {
  len_oj <- df$len[df$supp == "OJ" & df$dose == ds]
  len_vc <- df$len[df$supp == "VC" & df$dose == ds]

  test <- t.test(len_oj, len_vc,
                 paired = FALSE, var.equal = FALSE)

  tidy_test <- test %>%
    tidy %>%
    mutate(
      H_0 = paste0("mu_OJ - mu_VC = 0"),
      H_a = paste0("mu_OJ - mu_VC > 0"),
      dose = paste0(ds, " mg"),
      t = statistic,
      df = parameter,
      est_mu_OJ = test$estimate[1],
      est_mu_VC = test$estimate[2])

  # tidy result row is appended to pairwise_tests
  pairwise_tests <- bind_rows(pairwise_tests, tidy_test)
}

pairwise_tests <- pairwise_tests %>%
  # Apply Benjamini-Hochberg correction to p-value for each test
  # Controls the false discovery rate (FDR)
  mutate(adjusted_p = p.adjust(p.value, method = "BH")) %>%
```

```r
  select(H_0, H_a, dose,
         t, df, p.value, adjusted_p,
         conf.low, conf.high,
         est_mu_OJ, est_mu_VC)

# print.data.frame(pairwise_tests)

final_result <- bind_rows(result_all, pairwise_tests)

print.data.frame(final_result)
```

Code for Dose Level t-tests

```r
dose_levels <- sort(unique(df$dose)) # Sorted in ascending order
comparisons <- combn(dose_levels, 2, simplify = TRUE)
n_comparisons <- ncol(comparisons) # Number of comparisons to be made

pairwise_tests <- data.frame()

for (i in 1 : n_comparisons) {
  dose_x <- comparisons[2, i] # Take the larger dose first
  dose_y <- comparisons[1, i] # and the smaller dose second

  test <- t.test(df$len[df$dose == dose_x], df$len[df$dose == dose_y],
                 paired = FALSE, var.equal = FALSE)

  tidy_test <- test %>%
    tidy %>%
    mutate(
      # comparison = paste0("mu_", dose_x, " - mu_", dose_y),
      H_0 = paste0("mu_", dose_x, " - mu_", dose_y, " = 0"),
      # e.g. mu_1 - mu_0.5 = 0
      H_a = paste0("mu_", dose_x, " - mu_", dose_y, " != 0"),
      # e.g. mu_1 - mu_0.5 != 0
      t = statistic,
      df = parameter,
      est_mu_dose_x = test$estimate[1],
      est_mu_dose_y = test$estimate[2])

  # tidy result row is appended to pairwise_tests
  pairwise_tests <- bind_rows(pairwise_tests, tidy_test)
}

pairwise_tests <- pairwise_tests %>%
  # Apply Benjamini-Hochberg correction to p-value for each test
  # Controls the false discovery rate (FDR)
  mutate(adjusted_p = p.adjust(p.value, method = "BH")) %>%
  select(H_0, H_a,
         t, df, p.value, adjusted_p,
         conf.low, conf.high,
         est_mu_dose_x, est_mu_dose_y)

print.data.frame(pairwise_tests)
```