The effect of vitamin C on tooth growth in guinea pigs

Student

August 19, 2015

Introduction

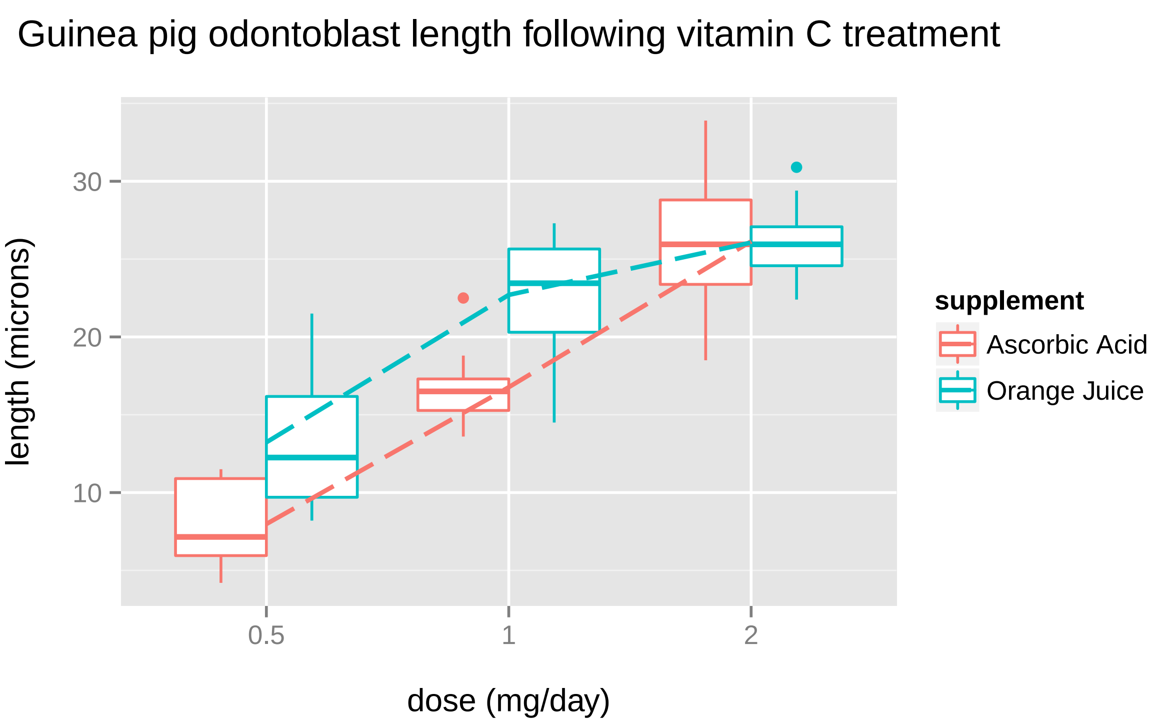
In this assignment we'll use **R** to explore, study, and draw conclusions about the effect of vtamin C on tooth growth in guinea pigs.

The data used is the **ToothGrowth** dataset from R's `datasets` package, and contains observations from 60 guinea pigs, divided into six equal groups. Each group of 10 guinea pigs received one of three dosages of vitamin C (0.5, 1, and 2 mg/day) via one of two delivery methods (orange juice or ascorbic acid), and post-treatment odontoblast length was measured.

The data comes from a [1947 study by Crampton, E.W](#).

Exploratory Data Analysis

To visually explore the sample data from each of the six groups of 10 guinea pigs, we use a boxplot. The boxplot shows quartiles and outliers (indicated by solid dots). Additionally, we've drawn a dashed line connecting the mean of each group. Generating code may be found in the appendix.



A few things are apparent about the *sample* data; odontoblast length were on average higher in groups with higher dosage, and at `0.5 mg/day` and `1.0 mg/day` dosage groups, supplement means were quite distinct. However, in order to make inferences about *population* means, we turn to hypothesis testing.

Hypothesis testing

We will test for differences in the unknown population means between various sets of two guinea pig groups. For any two populations for which we have a sample X and Y, we construct the following hypotheses:

1. **Null hypothesis** H_0: $\mu_X - \mu_Y = 0$
2. **Alternate hypothesis** H_a: $\mu_X - \mu_Y \neq 0$

It is assumed that X and Y are i.i.d normal random variables, with sample means $\bar{X}$ and $\bar{Y}$ following a known distribution. Because each sample group has small size $n = 10$, we treat their sample means as following a t, rather than normal distribution.

All 60 guinea pigs in the dataset are unique (see R datasets documentation, and so for any two groups we assume X and Y to be independent of one another. No two groups involve paired subjects.

The box plot reveals significant differences in sample variance between groups, and in such cases we opt for **Welch's *t* test**, a two-sample independent t test which does not assume equal population variances.

Conducting the tests

We will use **R**'s `t.test` function in a custom wrapper for terse output. The **ToothGrowth** dataset is supplied to the `tg` variable as a data.table. Full code is supplied in the Appendix.

In each case, we will construct H_0 and H_a as above.

Our wrapper function `welch.test` will display the 95% confidence interval, as well as the `p value`. For $p \leq 0.5$ we may express high confidence in preferring the alternate hypothesis, as a sample observation so extreme would be unlikely under the null hypothesis. The confidence interval will let us know whether we may say X or Y is expected to be greater.

Differences in dosage

Where X is the lower dosage group and Y is the higher dosage group, we evaluate for each supplement type:

Orange Juice

```
welch.test(length ~ dose,
  data = tg[(dose == 0.5 | dose == 1.0) & (supplement == "Orange Juice")])
```

```
## 95% confidence interval:  -13.4156, -5.52437 p-value:  8.78492e-05
```

```
welch.test(length ~ dose,
  data = tg[(dose == 1.0 | dose == 2.0) & (supplement == "Orange Juice")])
```

```
## 95% confidence interval:  -6.53144, -0.188557 p-value:  0.0391951
```

```
welch.test(length ~ dose,
  data = tg[(dose == 0.5 | dose == 2.0) & (supplement == "Orange Juice")])
```

```
## 95% confidence interval:  -16.3352, -9.32476 p-value:  1.32378e-06
```

Ascorbic Acid

```r
welch.test(length ~ dose,
  data = tg[(dose == 0.5 | dose == 1.0) & (supplement == "Ascorbic Acid")])
```

```
## 95% confidence interval:  -11.2657, -6.31429 p-value:  6.81102e-07
```

```r
welch.test(length ~ dose,
  data = tg[(dose == 1.0 | dose == 2.0) & (supplement == "Ascorbic Acid")])
```

```
## 95% confidence interval:  -13.0543, -5.68573 p-value:  9.1556e-05
```

```r
welch.test(length ~ dose,
  data = tg[(dose == 0.5 | dose == 2.0) & (supplement == "Ascorbic Acid")])
```

```
## 95% confidence interval:  -21.9015, -14.4185 p-value:  4.68158e-08
```

Differences in supplement

Where X is the Ascorbic Acid group and Y is the Orange Juice group, we evaluate for each dosage level:

```r
welch.test(length ~ supplement, data = tg[dose == 0.5])
```

```
## 95% confidence interval:  -8.78094, -1.71906 p-value:  0.00635861
```

```r
welch.test(length ~ supplement, data = tg[dose == 1.0])
```

```
## 95% confidence interval:  -9.05785, -2.80215 p-value:  0.00103838
```

```r
welch.test(length ~ supplement, data = tg[dose == 2.0])
```

```
## 95% confidence interval:  -3.63807, 3.79807 p-value:  0.963852
```

Conclusions

For both Orange Juice and Ascorbic Acid treated populations, there is statistically significant ($p \leq 0.5$) support for greater mean odontoblast length with increase in dosage, as shown by strictly positive 95% confidence intervals.

At `0.5` and `1.0` mg/day dosage, there is statistically significant ($p \leq 0.5$) support for greater opulation mean in the Orange Juice versus the Ascorbic Acid treated populations, as shown by the 95% confidence intervals.

At `2.0` mg/day dosage, the remarkably high p value ($1 - p \leq 0.05$) expresses high confidence in the null hypothesis. We must consider it likely for the population means between the Orange Juice and Ascorbic Acid treated populations to be equivalent.

Appendix

Confidence intervals

Welch's t test assigns a confidence interval for the difference in population means $\mu_X - \mu_Y$ in terms of known or calculable sample statistics. For a chosen $(1-\alpha)\cdot 100\%$ confidence level (usually 95%), it is given by:

$$\bar{X} - \bar{Y} \pm t_{\alpha/2,df}\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$$

where t is the critical t value (calculable via `qt()`) and the degrees of freedom df is given by:

$$df = \frac{(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y})^2}{\frac{(s_X^2/n_X)^2}{n_X - 1} + \frac{(s_Y^2/n_Y)^2}{n_Y - 1}}$$

If we chose $\alpha = 0.5$, the resulting 95% confidence interval expresses the idea that 95% of samples will yield a confidence interval that contains the true statistic $\mu_X - \mu_Y$. Therefore, it's unlikely that our observed sample has produced a calculated confidence interval that does *not* include $\mu_X - \mu_Y$, and more simply, we may say there is 95% confidence that the true statistic falls within the interval observed.

The immediate consequences of the 95% confidence interval are:

1. If the interval contains zero, then there is not enough evidence to confidently assert that H_0 is unlikely.
2. If the interval does not contain zero, then H_0 is considered unlikely by the evidence.
   a. If the interval is strictly positive, then we express confidence that $\mu_X > \mu_Y$.

   b. If the interval is strictly negative, then we express confidence that $\mu_Y > \mu_X$.

The p value

The p value is the result of the statistical test which we select, and quantifies the success or failure of the null hypothesis H_0. It is the probability that the null hypothesis version of population parameters could have yielded the observed sample statistic.

If the p value is very small, then the evidence against the null hypothesis is considered to be great. Usually we consider this to be for $p \leq 0.05$. In other words, it is considered unlikely to have observed so extreme a result as we have in the sample.

Variance comparison

To complement the boxplot, we can also calculate sample variance for each group explicitly. Note that between some groups, the variances may be regarded as reasonably close, but it seems more reasonable to avoid assuming that guinea pigs with either different vitamin C intake methods or dosage levels would have identical population variance in odontoblast length. Welch's t test is thus more generally applicable, and it is the test conducted by default via `t.test`.

```
tg[, .(variance = var(length)), by = .(supplement, dose)]
```

```
##       supplement dose  variance
## 1: Ascorbic Acid  0.5  7.544000
## 2: Ascorbic Acid  1.0  6.326778
## 3: Ascorbic Acid  2.0 23.018222
## 4:  Orange Juice  0.5 19.889000
## 5:  Orange Juice  1.0 15.295556
## 6:  Orange Juice  2.0  7.049333
```

Preliminary R code

```r
# Libraries.
library(data.table)
library(ggplot2)

# Load and process data.
data(ToothGrowth)
tg <- data.table(
  supplement = factor(
    ToothGrowth$supp,
    levels = unique(ToothGrowth$supp),
    labels = c("Ascorbic Acid", "Orange Juice")
  ),
  length = ToothGrowth$len,
  dose = ToothGrowth$dose
)

# ggplot settings
update_geom_defaults("point", list(colour = NULL))
```

Welch test wrapper function

```r
welch.test <- function(x, data, ...) {
  o <- t.test(x, data = data,
    alternative = "two.sided",
    mu = 0,
    paired = FALSE,
    var.equal = FALSE,
    conf.level = 0.95,
    ...
  )

  display.conf.int <- paste(
    signif(o$conf.int[1], 6),
    signif(o$conf.int[2], 6),
    sep = ", "
  )

  cat("95% confidence interval: ", display.conf.int,
      "p-value: ", signif(o$p.value, 6))
}
```

Plot code

```
ggplot(tg, aes(x = factor(dose), y = length, colour = supplement)) +
  geom_boxplot() +
  stat_summary(
    fun.y = mean, aes(group = supplement),
    geom = "line",
    linetype = "longdash",
    size = 0.8
  ) +
  labs(x = "\ndose (mg/day)", y = "length (microns)\n") +
  labs(title = "Guinea pig odontoblast length following vitamin C treatment\n")
```

Sample verbose t.test output

```
t.test(length ~ supplement, data = tg[dose == 0.5])
```

```
##
##  Welch Two Sample t-test
##
## data:  length by supplement
## t = -3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.780943 -1.719057
## sample estimates:
## mean in group Ascorbic Acid  mean in group Orange Juice
##                        7.98                       13.23
```

```
t.test(length ~ dose,
  data = tg[(dose == 0.5 | dose == 1.0) & supplement == "Orange Juice"])
```

```
##
##  Welch Two Sample t-test
##
## data:  length by dose
## t = -5.0486, df = 17.698, p-value = 8.785e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13.415634  -5.524366
## sample estimates:
## mean in group 0.5   mean in group 1
##             13.23             22.70
```

Closing Remarks

Thank you for reading! Hope you are enjoying the class(es)!