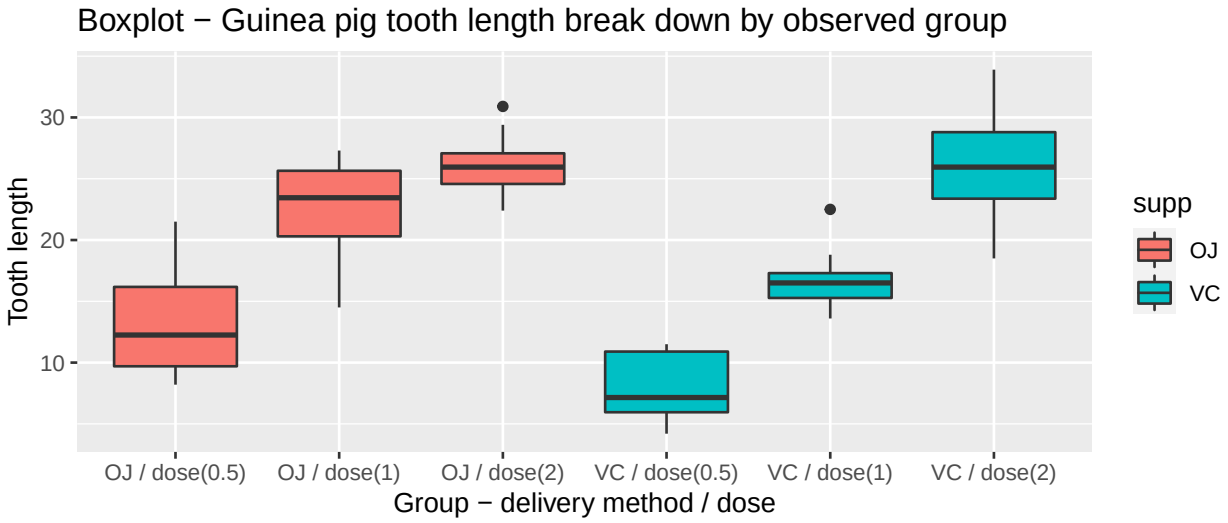
Statistical Inference Course Project

Marko Intihar

Part 2: Basic Inferential Data Analysis

In part two of the project we will analyze the **"ToothGrowth"** data in the R **datasets** package. Data holds results for the effect of Vitamin C on Tooth Growth in Guinea Pigs. Observed variable is X-tooth length of Guinea pig. 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 (first group **OC**) or ascorbic acid (a form of vitamin C and coded as group **VC**). We have 2 delivery methods times 3 different doses resulting in 6 different subgroups of the data. Let's draw a boxplot for observed variable for each group:

```
ToothGrowth %>%
  mutate(group = paste(supp, " / dose(", dose, ")", sep = "")) %>%
  ggplot(aes(x = group, y = len, fill = supp)) + geom_boxplot() +
  labs(title = "Boxplot - Guinea pig tooth length break down by observed group",
       x = "Group - delivery method / dose", y = "Tooth length")
```



We have observed total of 60 Guinea pigs, each group consists of $n = 10$ animals. As we can see from the provided box plot above, there are differences in observed tooth length for given groups. If we increase dose the tooth length increases for each delivery method. Also we can see there are differences between delivery methods for given dose. But we must stress that we have relative small sample sizes for each group, so results may vary.

For our formal statistical test (hypothesis testing) we would like to check if given sample data indicate there are differences in mean tooth length (for population of Guinea pigs) if delivery method is used "OJ" (group 1) or "VC" (group 2) with the same dosage 1 mg/day. In order to do proper testing some assumptions must be made: 1) Random variable ($\bar{X}$) sample tooth mean follows normal distribution (CLT) | 2) We do not know variance of X for both groups of populations so we will assume we have different variances (to be on the safe side!) | 3) we are conducting two groups t-test for means (we do not have paired samples!) | 4) Selected α level is 5%. | 5) Our hypothesis are:

- $H_0 : \mu_1 = \mu_2$ (populations means are equal)
- $H_a : \mu_1 \neq \mu_2$ (populations means are different)

6) We are using two sided t test. Lets conduct a test:

```
# conduct a test
test <- t.test(x = ToothGrowth %>% filter(supp == "OJ" & dose == 1) %>% pull(len),
               y = ToothGrowth %>% filter(supp == "VC" & dose == 1) %>% pull(len),
               alternative = "two.sided", paired = FALSE, var.equal = FALSE)
test[c(5,4,3)]
```

```
## $estimate
## mean of x mean of y
##     22.70     16.77
##
## $conf.int
## [1] 2.802148 9.057852
## attr(,"conf.level")
## [1] 0.95
##
## $p.value
## [1] 0.001038376
```

R's test results returned sample means, where first group sample mean is 22.7 and second group sample mean is 16.77. This indicates there are quite big differences in both sample means. Also mean difference confidence interval conducted at at significance level $\alpha = 5\%$ shows that value 0 is not included in the interval, which indicates that there are differences in population means for selected two groups. Rejection of null hypothesis H_0 is formally confirmed with p-value 0.0010384, which is lower than $\alpha = 5\%$. Therefore we can reject H_0 at 95% confidence level, and we can say that mean tooth length of Guinea pig is different for both selected groups (population level), here for the first group method used was "OJ" (dosage 1 mg/day) and for the second group method used was "VC" (dosage 1 mg/day).