ToothGrowthAnalyses

AlekseyVosk

3 June 2018

For research we will need functions from external packages.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------------

## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.4
## v tidyr   0.8.0     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

## -- Conflicts ------------------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Data and exploratory data analyses

Dataset contanes 60 observations on 3 variables:
1. len - Tooth length
2. supp - Supplement type: *VC* (ascorbic acid) or *OJ* (orange juice)
3. dose - Dose in milligrams/day

Loading data as tibble and looking at data's structure.

```
dataset <- as_tibble(ToothGrowth)
str(dataset)
```

```
## Classes 'tbl_df', 'tbl' and '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 ...
```

Let's compare tooth length in groups with different doses and different supplement types.

```
dataset_f <- mutate(dataset, dose = as.factor(dose))
table(dataset_f$supp, dataset_f$dose)
```

```
##
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

```
group_by(dataset_f, dose) %>%
    summarise(len = mean(len))
```

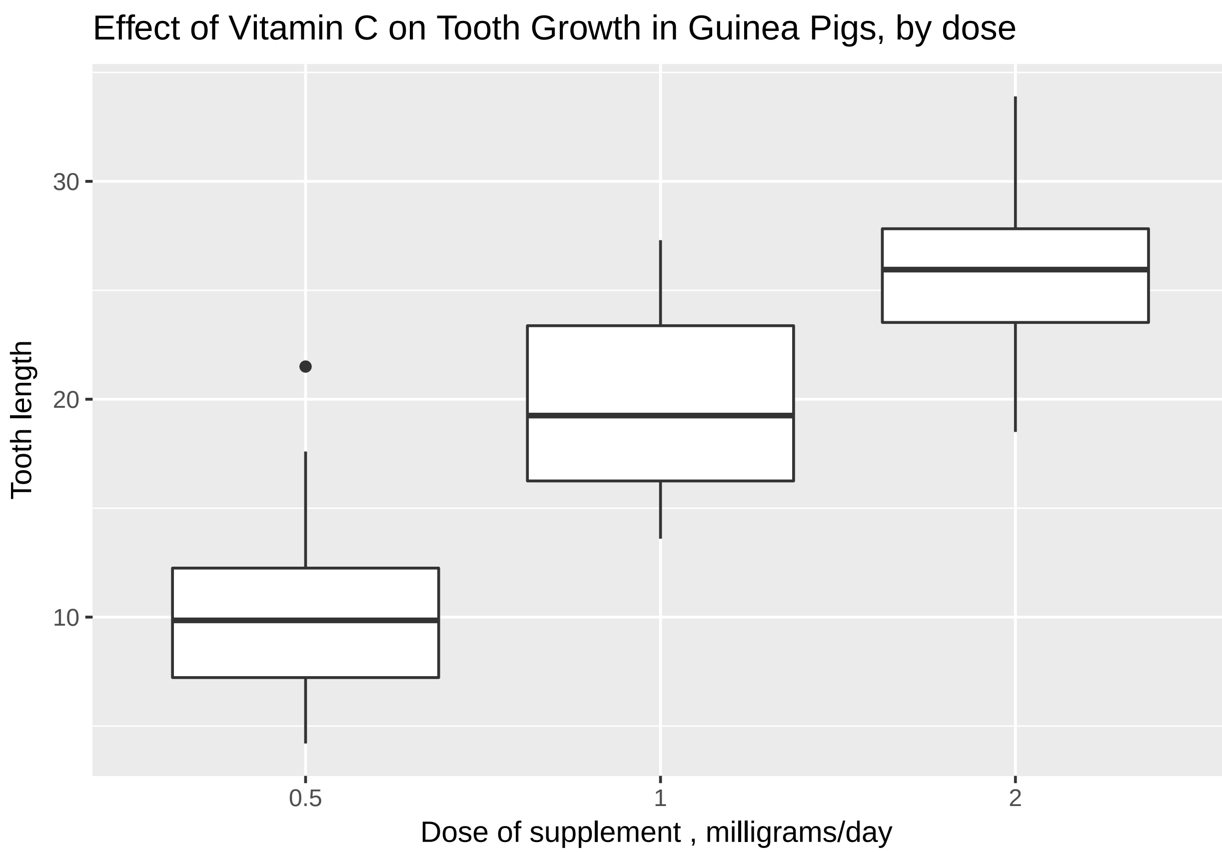
```
## # A tibble: 3 x 2
##   dose    len
##   <fct> <dbl>
## 1 0.5    10.6
## 2 1      19.7
## 3 2      26.1
```

```r
group_by(dataset_f, supp) %>%
    summarise(len = mean(len))
```

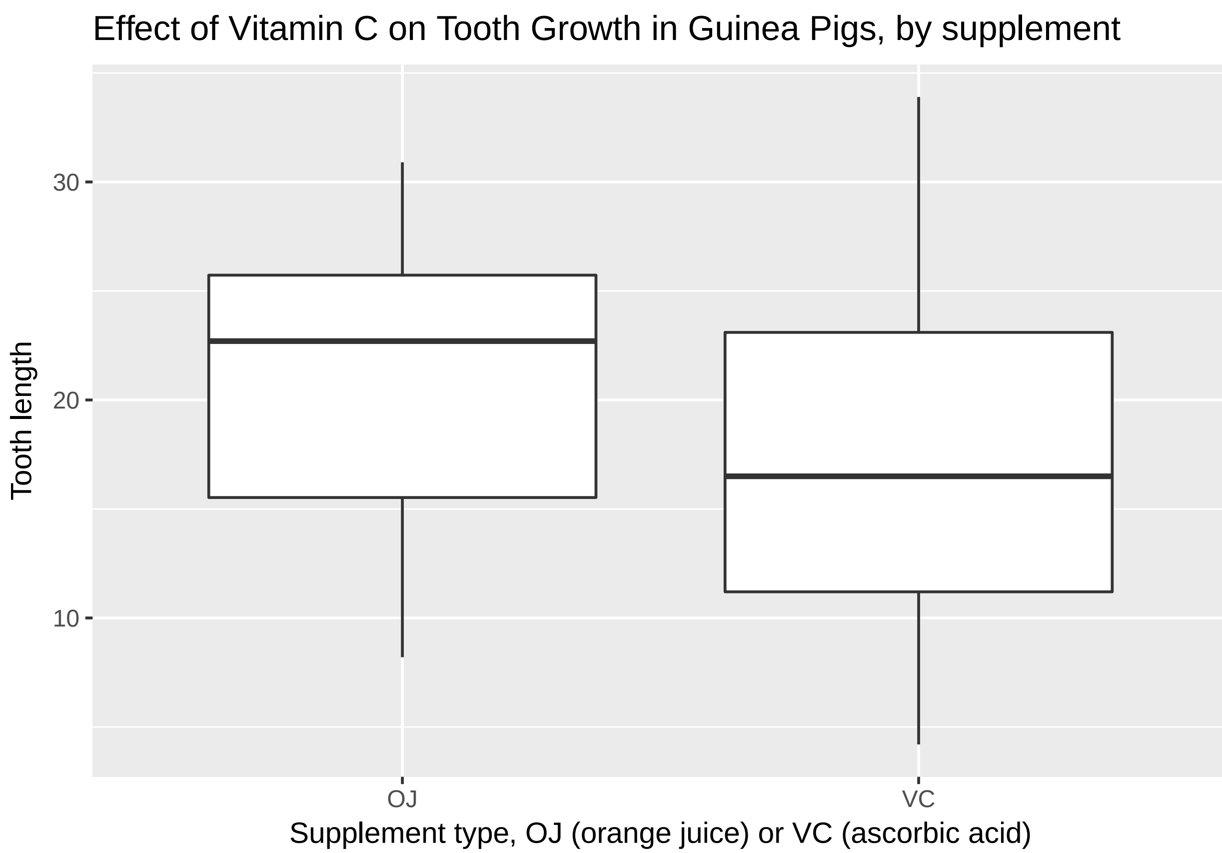
```
## # A tibble: 2 x 2
##   supp    len
##   <fct> <dbl>
## 1 OJ     20.7
## 2 VC     17.0
```

Tooth length differs between each groups, to visualise it, let's make boxplots:

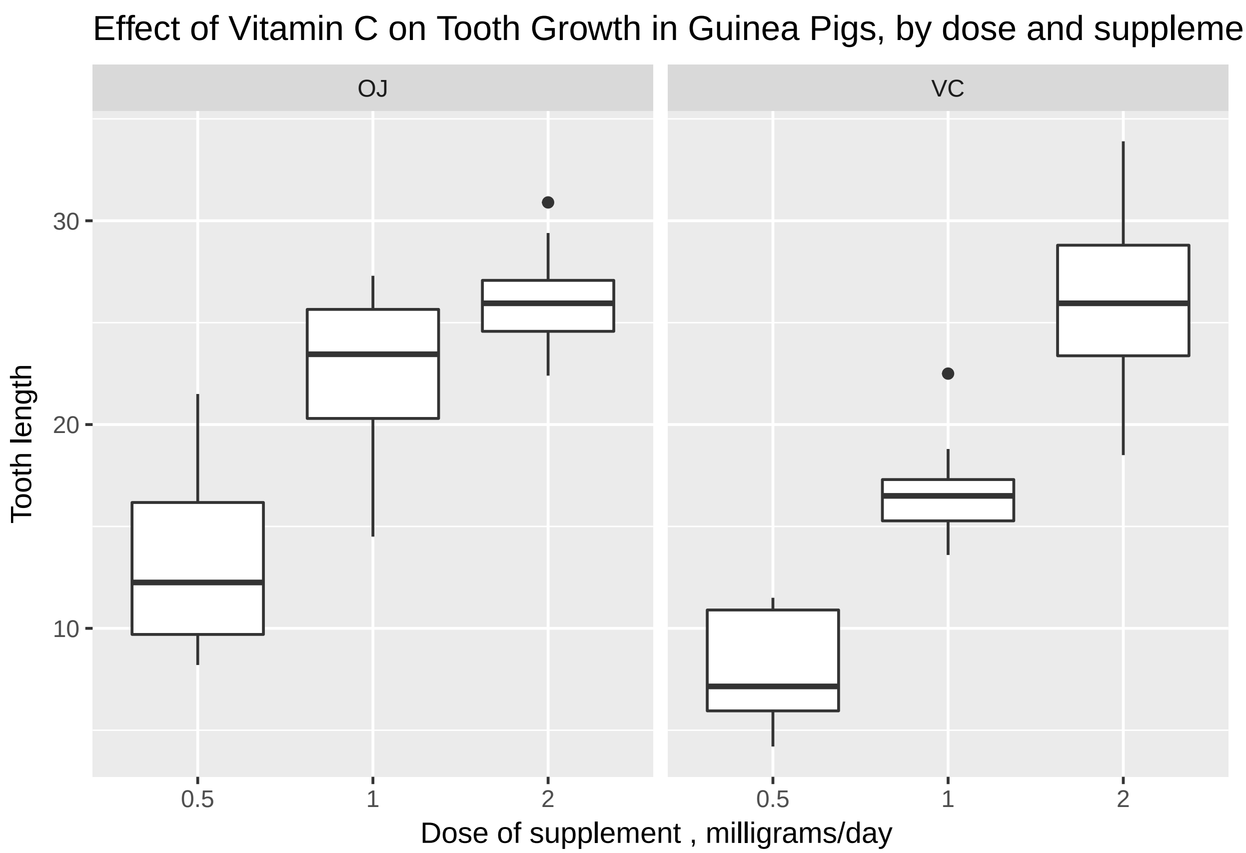
```r
g <- ggplot(data = dataset_f, mapping = aes(y = len)) +
    labs(y = "Tooth length",
         title = "Effect of Vitamin C on Tooth Growth in Guinea Pigs")
g + geom_boxplot(mapping = aes(x = dose)) +
    labs(x = "Dose of supplement , milligrams/day ",
         title = "Effect of Vitamin C on Tooth Growth in Guinea Pigs, by dose")
```




```r
g + geom_boxplot(mapping = aes(x = supp)) +
    labs(x = "Supplement type, OJ (orange juice) or VC (ascorbic acid)",
         title = "Effect of Vitamin C on Tooth Growth in Guinea Pigs, by supplement")
```




```
g + geom_boxplot(mapping = aes(x = dose)) +
    facet_wrap(~ supp) +
    labs(x = "Dose of supplement , milligrams/day ",
         title = "Effect of Vitamin C on Tooth Growth in Guinea Pigs, by dose and supplement")
```



Exploratory analyses shows us, than the tooth growth in guinea pigs differs in groups with different doses of vitamin C, but not so much in groups with different supplement type.

Statistical inference

For statistical inference first estimate relationships between tooth length and supplement types

Hypotheses tests for tooth length and supplement types:
H0: Tooth length and supplement types are **independent**. Tooth length dictribution **does not vary** by the supplement types.
HA: Tooth length and supplement types are **dependent**. Tooth length dictribution **vary** by the supplement types.

We compare one numerical and one categorical (with 2 levels and 30 cases in each category) variables, so I'll use Student's t-Test with significance level 0.05.

```
test_len_supp <- t.test(filter(dataset, supp == "VC")$len,
                        filter(dataset, supp == "OJ")$len,
                        alternative = "two.sided",
                        mu = 0,
                        paired = FALSE,
                        conf.level = 0.95)
paste0("p-value: ",test_len_supp$p.value)
```

```
## [1] "p-value: 0.0606345078809341"
```

P-value is above significance level, so we fail to regect H0.
We can also see 95% confidence interval:

```
test_len_supp$conf.int
```

```
## [1] -7.5710156  0.1710156
## attr(,"conf.level")
## [1] 0.95
```

Confidence interval includes zero, so we can't rule out possibility of indifference between 2 groups.

Now let's estimate relationships between tooth length and dose of vitamin C

Hypotheses tests for tooth length and dose:
H0: Tooth length and vitamin dose are **independent**. Tooth length dictribution **does not vary** by the vitamin dose.
HA: Tooth length and vitamin dose are **dependent**. Tooth length dictribution **vary** by the vitamin dose.

We compare one numerical and one categorical (with 3 levels and 20 cases in each category) variables. The optimal statistical test would be ANOVA, but as we need to use technique from the class, let's conduct 3 Student's t-Tests (still 2-sided).
We will need to correct our significance level for this 3 test's, for simplicity let's make Bonferroni correction on 0.05:

```
paste0("Bonferroni significance level: ", 0.05/3)
```

```
## [1] "Bonferroni significance level: 0.0166666666666667"
```

```
test0.5_1 <- t.test(filter(dataset, dose == 0.5)$len,
                    filter(dataset, dose == 1)$len)
test1_2 <- t.test(filter(dataset, dose == 1)$len,
                    filter(dataset, dose == 2)$len)
test0.5_2 <- t.test(filter(dataset, dose == 0.5)$len,
                    filter(dataset, dose == 2)$len)
paste0("0.5/1 p-value: ",test0.5_1$p.value)
```

```
## [1] "0.5/1 p-value: 1.26830072017385e-07"
```

```
paste0("1/2 p-value: ",test1_2$p.value)
```

```
## [1] "1/2 p-value: 1.9064295136718e-05"
```

```
paste0("0.5/2 p-value: ",test0.5_2$p.value)
```

```
## [1] "0.5/2 p-value: 4.39752495936323e-14"
```

All 3 p-values are much lower, then corrected significance level, so we can reject H0 in favor of HA.

Conclusions

Our statistical inference shows that at usual significance level tooth length and vitamin doses are associated, but tooth length and supplement types are not.
For this conclusions we must assume, that all pigs were independent to each other, and each pig corresponds to only one observation in the dataset.