# Statistical Inference Course Project - Part II

*Steven Michiels*

*4/17/2020*

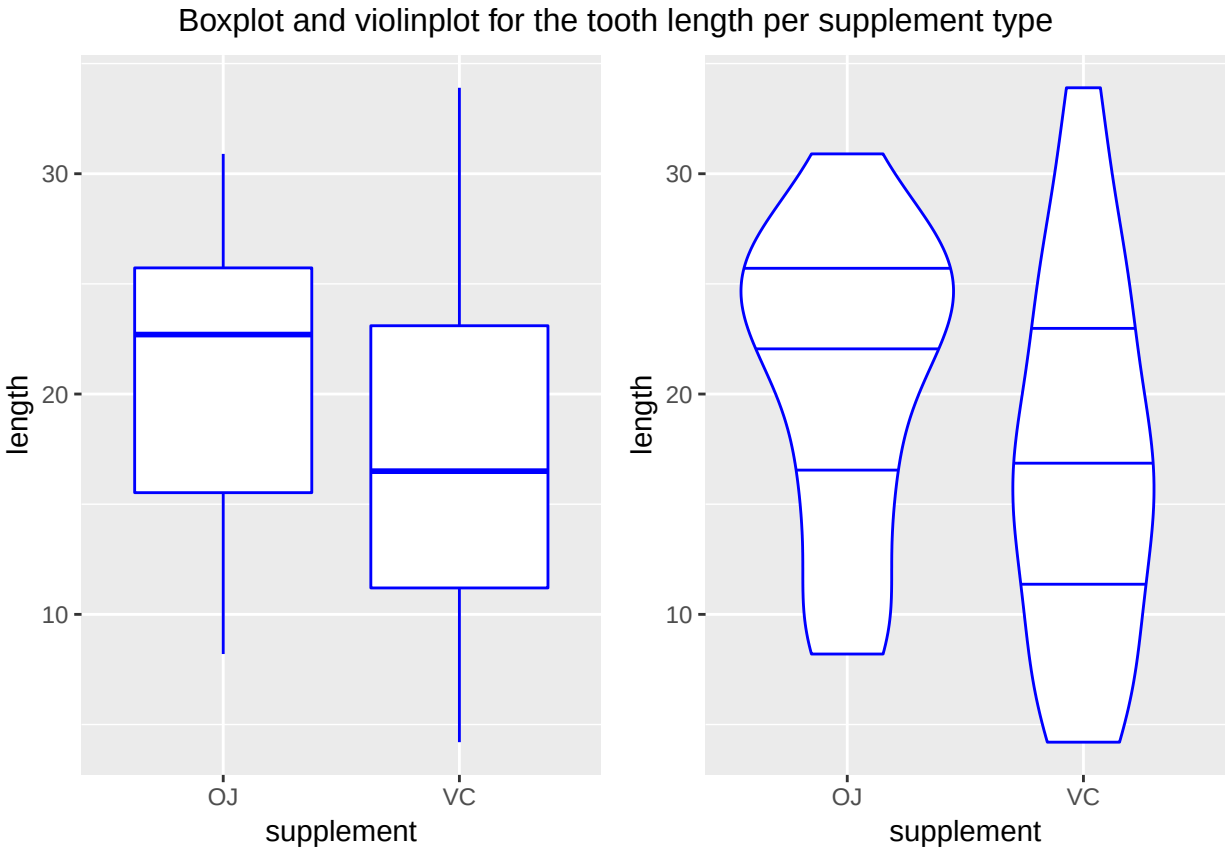
## Exploratory data analysis

We load the data and convert the dose predictor to factors.

```
## [1] "len"  "supp" "dose"
```

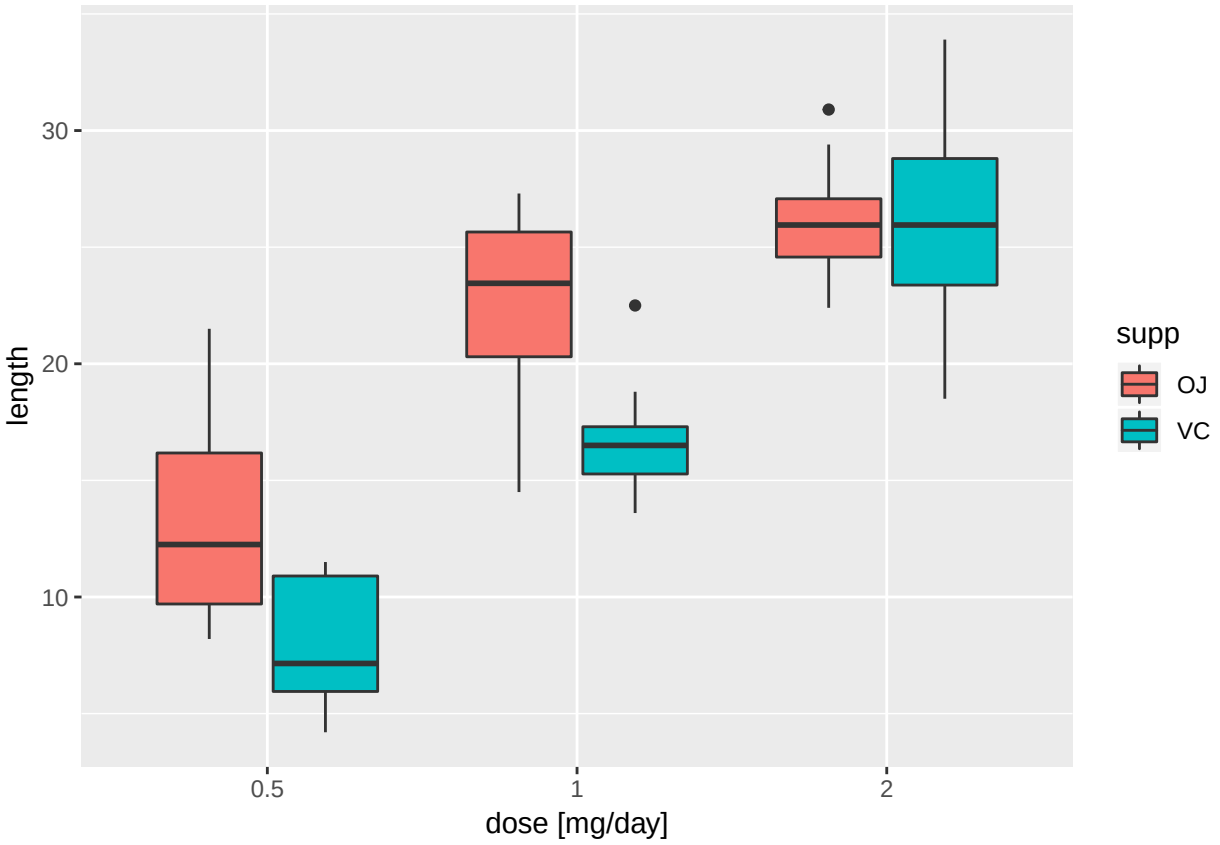
We create a **boxplot and violinplot** for the **tooth length in function of the used supplement type**. The **OJ** supplement **appears** to be **associated with a higher tooth length**, although statistical significance is yet to be shown. The **OJ supplement type** distribution appears to be **somewhat skewed**.

```
g1a=ggplot(ToothGrowth,aes(x = supp,y=len))
g1a=g1a + geom_violin(color="blue", fill="white", draw_quantiles = c(0.25, 0.5, 0.75),scale = "count")
g1a=g1a+xlab("supplement")+ylab("length")

g1b=ggplot(ToothGrowth,aes(x = supp,y=len))
g1b=g1b + geom_boxplot(color="blue", fill="white")
g1b=g1b+xlab("supplement")+ylab("length")

grid.arrange(g1b,g1a,nrow=1,top="Boxplot and violinplot for the tooth length per supplement type")
```



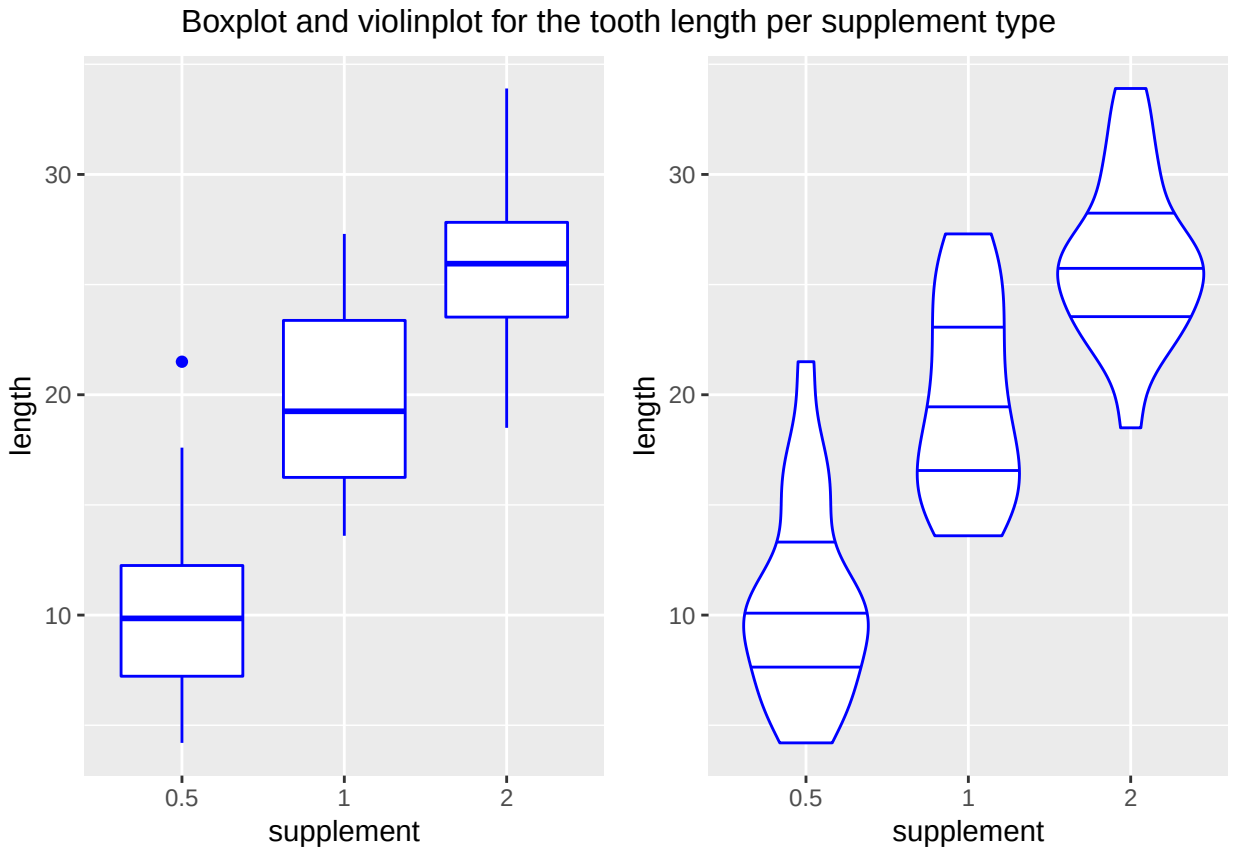
We create a **boxplot** for the **tooth length in function of the supplement**, but now we use **supgroups per dose level**. We now see that the **association of tooth length with the supplement type may only be present for the two lower dose levels**, i.e. .5 mg/day and 1.0 mg/day.

```
g2=ggplot(ToothGrowth,aes(x = factor(dose),y=len, fill=supp))
g2=g2 + geom_boxplot()
#g=g+geom_smooth(method="lm", formula=y~x)
g2=g2+xlab("dose [mg/day]")+ylab("length")
plot(g2)
```



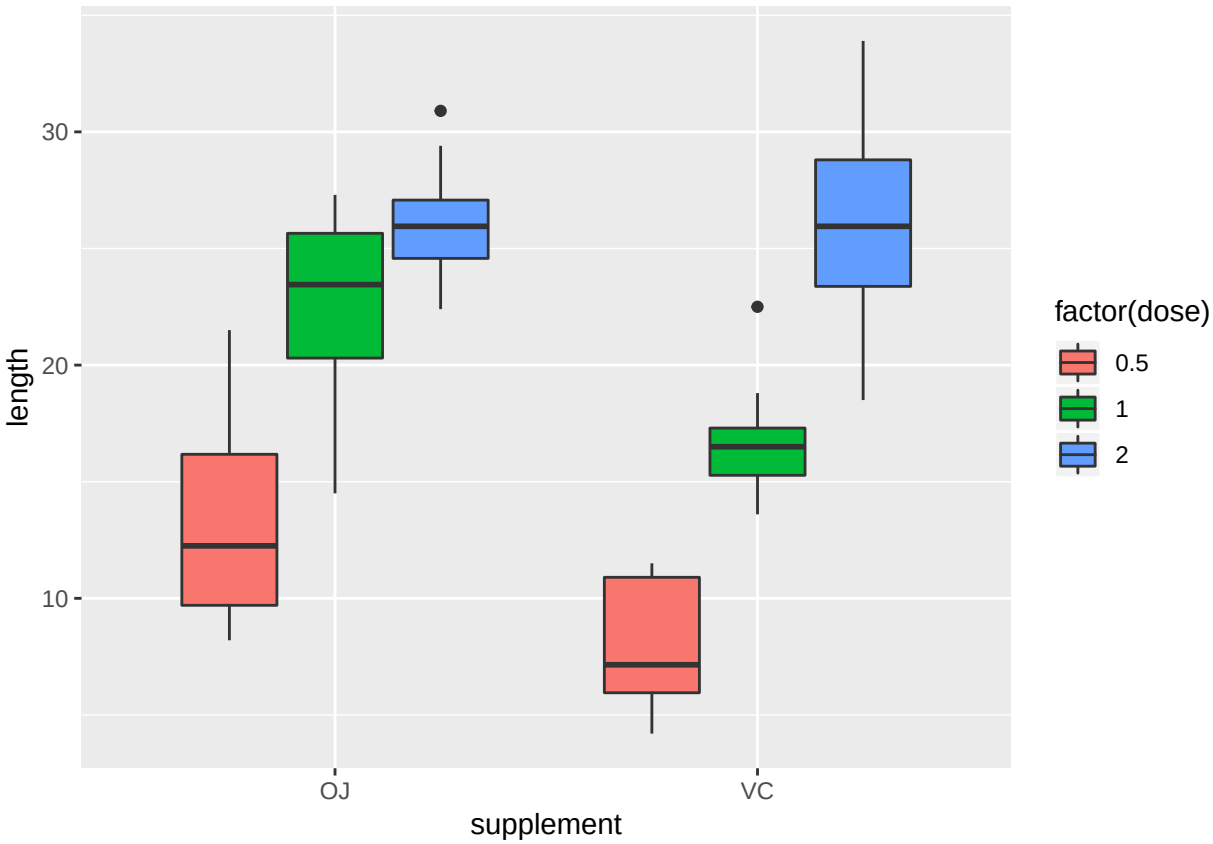

We create a **boxplot and violinplot** for the **tooth length in function of the used dose**. **Increased dose** of the supplement **appears** to be associated with an **increased tooth length**.

```
g3a=ggplot(ToothGrowth,aes(x = factor(dose),y=len))
g3a=g3a + geom_violin(color="blue", fill="white", draw_quantiles = c(0.25, 0.5, 0.75),scale = "count")
g3a=g3a+xlab("supplement")+ylab("length")

g3b=ggplot(ToothGrowth,aes(x = factor(dose),y=len))
g3b=g3b + geom_boxplot(color="blue", fill="white")
g3b=g3b+xlab("supplement")+ylab("length")

grid.arrange(g3b,g3a,nrow=1,top="Boxplot and violinplot for the tooth length per supplement type")
```




We create a **boxplot** for the ___tooth length in function of the used dose, but now we make **subgroups per supplement type**. We now see a rather **clear association between the dose level and the tooth length**, although this association appears to be weaker at the higher dose levels for the OJ supplement type.

```
g3=ggplot(ToothGrowth,aes(x = supp,y=len,fill=factor(dose)) )
g3=g3 + geom_boxplot()
g3=g3+xlab("supplement")+ylab("length")
plot(g3)
```




## Basic summary of the data.

```
summary(ToothGrowth)
```

```
##       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
```

## Hypothesis testing

We create different subgroups to perform different subgroup analyses.

First, we perform a two-sided t-test of a difference in mean tooth length between the different dose levels. The **tooth length** is **clearly associated with the dose level**: at a **significance level of 5%**, the mean tooth length is different between all the dose levels.

```
t_dose05_10=t.test(dose05$len,dose10$len)
t_dose10_20=t.test(dose10$len,dose20$len)
t_dose05_20=t.test(dose05$len,dose20$len)
t_dose05_10$p.value
```

```
## [1] 1.268301e-07
```

```
t_dose10_20$p.value
```

```
## [1] 1.90643e-05
```

```
t_dose05_20$p.value
```

```
## [1] 4.397525e-14
```

Next, we perform a two-sided t-test of a difference in mean tooth length between the two supplement types. The **tooth length for supp OJ** shows a **trend towards a higher mean** than for supp VC, with a p-value only slightly above a 5% significance level and with a 95* CI barely including zero as the difference in mean. Nevertheless, at a **significance level of 5%** we still **fail to reject the null hypothesis of equal mean length**.

```
t_supp=t.test(suppOJ$len,suppVC$len)
t_supp
```

```
##
##  Welch Two Sample t-test
##
## data:  suppOJ$len and suppVC$len
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

Finally, we perform a two-sided t-test of a difference in mean tooth length between the supplement types but at each dose level separately. For the **lower dose levels 0.5 mg/day and 1.0 mg/day**, we find an **influence of the supplement type on the mean tooth length**, at a significance level of 5%, **but this influence disappears at the highest dose level**.

```
t_dose05_supp=t.test(dose05VC$len,dose05OJ$len)
t_dose10_supp=t.test(dose10VC$len,dose10OJ$len)
t_dose20_supp=t.test(dose20VC$len,dose20OJ$len)
t_dose05_supp$p.value
```

```
## [1] 0.006358607
```

```
t_dose10_supp$p.value
```

```
## [1] 0.001038376
```

```
t_dose20_supp$p.value
```

```
## [1] 0.9638516
```

# Conclusions and assumptions

## Conclusion

- The dose level has a significant effect on the tooth length.
- The supplement type alone does not affect the tooth length. Only at the two lowest dose levels, the supplement type has a significant effect on the tooth length.

## Assumptions

- Independent and identically distributed samples
- Unskewed, mount-shaped distributions (requirement for t-test)