Statistical Inference

Dorian Kcira

10/24/2015

Tooth Growth Data

The dataset ToothGrowth is loaded and a simple exploratory data analysis is done. From the help (?ToothGrowth): "The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid)."

We table the dose and supplement type. We also plot the dependency of the tooth length on the dose for the two delivery methods (OJ, VC).

```
library(lattice)
data(ToothGrowth)
str(ToothGrowth)
```

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

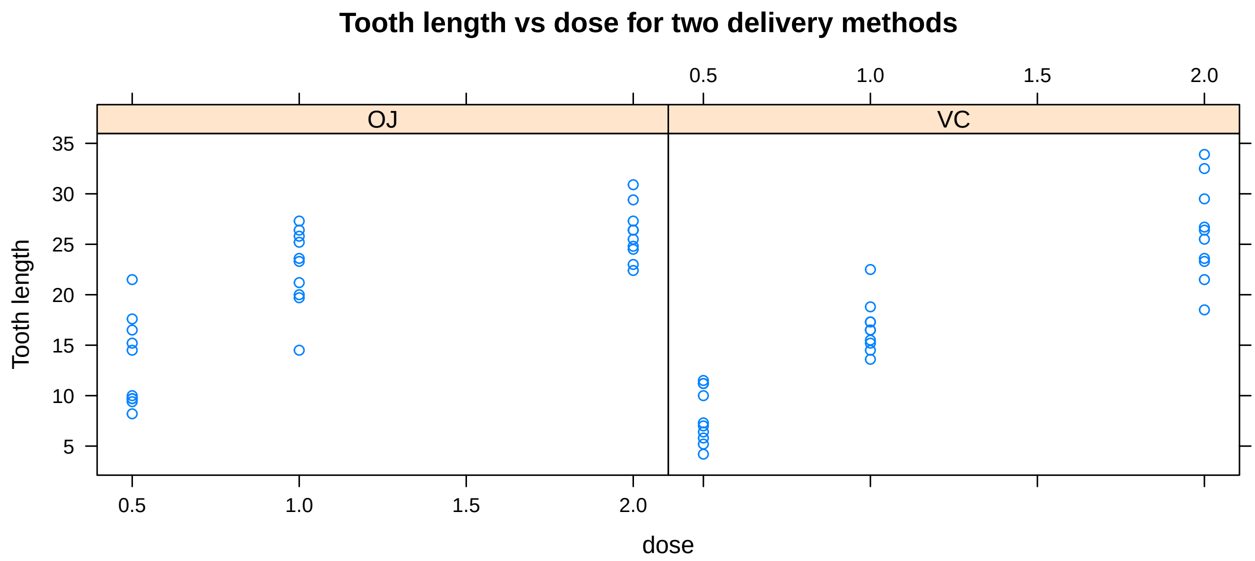
```
table(ToothGrowth$dose)
```

```
##
## 0.5   1   2
##  20  20  20
```

```
table(ToothGrowth$supp)
```

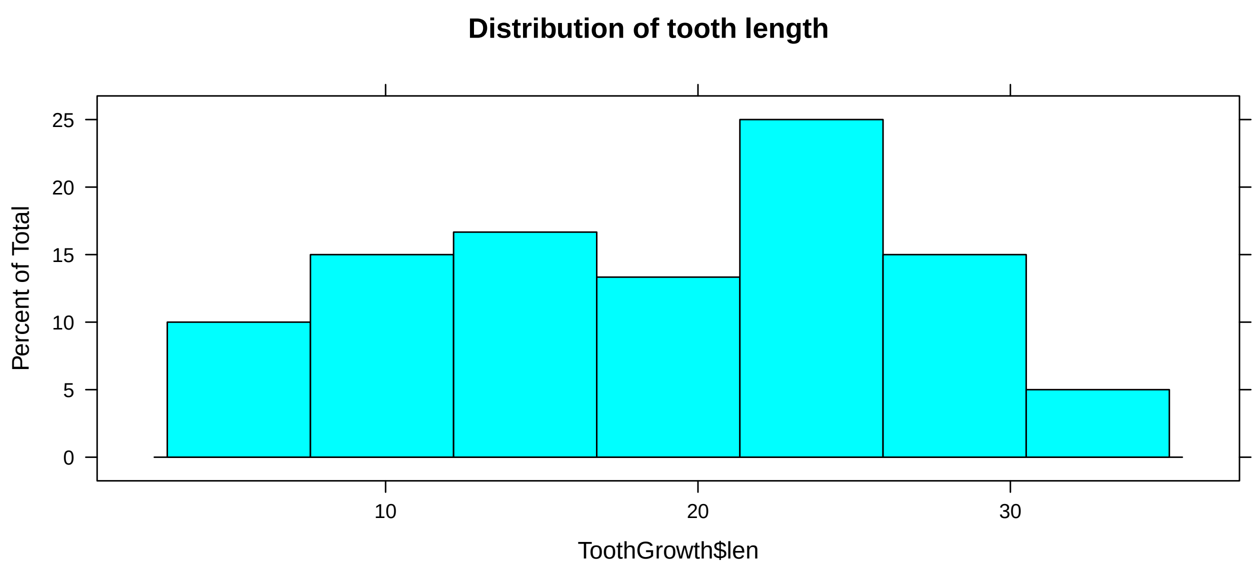
```
##
## OJ VC
## 30 30
```

```
with(data=ToothGrowth,
     xyplot(len~dose|supp,
       main="Tooth length vs dose for two delivery methods",
       ylab="Tooth length", xlab="dose")
)
```



We can see that there are large variations of the tooth length for each dose, but there is a clear tendency of teeth being longer for larger doses of vitamin C. Also, visually, the orange juice (OJ) method seems to do better than the absorbic acid (VC) method. We need to confirm this analytically.

```
histogram(ToothGrowth$len, main="Distribution of tooth length")
```



We also look at the histogram of the tooth length distribution to see if it has any particular shape.

Hypotheses

The problem formulation does not clarify whether the measurements were paired, therefore we will use the relaxed assumption that they are independent. From the panel plots above, the variance looks different among the different subsets. We will, therefore assume different variances for the t-tests below.

The other assumption will be that the tooth length data are normally distributed.

Our first hypothesis will be that the tooth length increases with the dose. Our second hypothesis will be that the tooth length depends on the delivery method.

Testing the Hypotheses

Tooth length depends on the vitamin C dose

The null hypothesis will be that the tooth length does not depend on the dose. Using the length data for each dose we build three t-tests for every combination and print out the 95% confidence intervals.

```
length05 = ToothGrowth$len[ToothGrowth$dose==0.5]
length1  = ToothGrowth$len[ToothGrowth$dose==1]
length2  = ToothGrowth$len[ToothGrowth$dose==2]
test1 <- t.test(length2 , length05 , paired = FALSE , var.equal = FALSE , conf.level = 0.95)
cat("The test1 95% confidence interval is:",test1$conf.int,"\n")
```

```
## The test1 95% confidence interval is: 12.83383 18.15617
```

```
test2 <- t.test(length2 , length1 , paired = FALSE , var.equal = FALSE , conf.level = 0.95)
cat("The test2 95% confidence interval is:",test2$conf.int,"\n")
```

```
## The test2 95% confidence interval is: 3.733519 8.996481
```

```
test3 <- t.test(length1 , length05 , paired = FALSE , var.equal = FALSE , conf.level = 0.95)
cat("The test3 95% confidence interval is:",test3$conf.int,"\n")
```

```
## The test3 95% confidence interval is: 6.276219 11.98378
```

In each case the confidence interval does not include the null hypothesis, i.e. we can conclude that tooth length depends on the vitamin C dose and increases with it.

Tooth length depends on the vitamin C delivery method

The null hypothesis here is that the tooth length does not depend on the delivery method.

```
lengthOJ = ToothGrowth$len[ToothGrowth$supp=="OJ"]
lengthVC = ToothGrowth$len[ToothGrowth$supp=="VC"]
testsupp = t.test(lengthOJ , lengthVC , paired = FALSE , var.equal = FALSE , conf.level = 0.95)
cat("The delivery method 95% confidence interval is:",testsupp$conf.int,"\n")
```

```
## The delivery method 95% confidence interval is: -0.1710156 7.571016
```

The confidence level here includes 0, hence we are not able to exclude the null hypothesis. So we are not able to prove that the tooth length depends on the delivery method. For the sake of completeness, let's re-test this last hypothesis with the assumption that the data are paired.

```
testsupp2 = t.test(lengthOJ , lengthVC , paired = TRUE , var.equal = FALSE , conf.level = 0.95)
cat("The delivery method for paired data 95% confidence interval is:",testsupp2$conf.int,"\n")
```

```
## The delivery method for paired data 95% confidence interval is: 1.408659 5.991341
```

If the data were paired, which we do not know for sure, we would be able to conclude that the tooth length also depends on the delivery method and it is higher for Orange Juice.

Conclusions

After doing a short exploratory analysis of the data we looked at the data with these assumptions:

1. the data (tooth length) are normally distributed
2. the measurements are independent for the different dose and delivery methods
3. the variance among the different doses and delivery methods are different

Our first hypothesis was that the tooth length increases with the dose. Using the t-test we demonstrated that this hypothesis is true and that the larger the dose, the larger the tooth length

Our second hypothesis was that the tooth length depends on the delivery method. Using the assumptions above we were not able to prove this hypothesis. If one changes the assumptions, assuming that the data are paired, this hypothesis is confirmed and the Orange Juice method has a larger impact on tooth length than ascorbic acid.