Basic inferential data analysis

Basic exploratory data analyses and summary

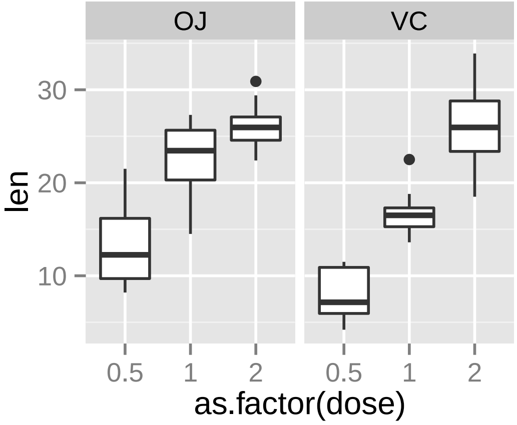
First, we load the data and change its name to a shorter one.

```
data(ToothGrowth)
tg <- ToothGrowth
rm(ToothGrowth)
str(tg)
```

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

We can see in the `str(tg)` output and the `help(ToothGrowth)` help page that there are 60 observations of 3 variables in the dataset. `tg$supp` and `tg$dose` can be treated as factors to see what are the differences between each delivery method and dose.

```
qplot(as.factor(dose), len, data = tg, facets = . ~ supp, geom="boxplot")
```



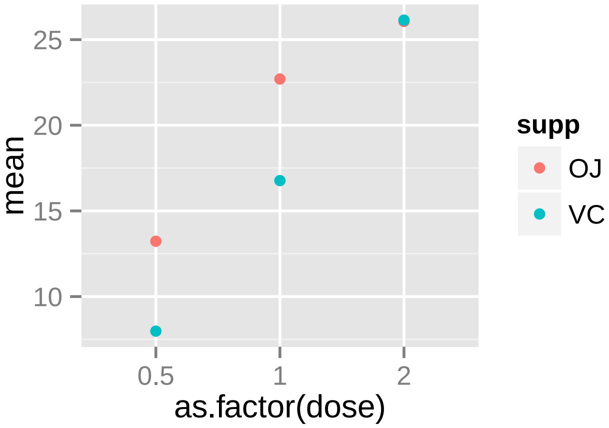
From the box plot we can see that, apparently, the orange juice (OJ) delivery method is better for 0.5 and 1 mg doses. The 2 mg dose is uncertain, since both methods seem to have the same median. The orange juice one, though, has less variability for this dose. Another interesting fact is that, when the doses were delivered in orange juices, the difference in length between the 1 and 2 mg doses are less than any other case. We should perform a t test between those values to see whether the 2 mg dose is significantly better than the 1 mg.

We can also calculate the mean and variance of the length per method and dose and plot the means.

```
bysuppdose <- ddply(tg, .(supp, dose), summarize, mean = mean(len), var = var(len))
bysuppdose
```

```
##   supp dose  mean    var
## 1   OJ  0.5 13.23 19.889
## 2   OJ  1.0 22.70 15.296
## 3   OJ  2.0 26.06  7.049
## 4   VC  0.5  7.98  7.544
## 5   VC  1.0 16.77  6.327
## 6   VC  2.0 26.14 23.018
```

```
qplot(as.factor(dose), mean, data = bysuppdose, colour = supp)
```



Again, there is evidence that the orange juice delivery method is better for the 0.5 and 1 mg dose, but the ascorbic acid growth mean is slightly better for the 2 mg dose.

Hypothesis testing

We can perform t-tests to test different hypothesis. First, we will test if the orange juice method is better for the 0.5 and 1 mg dose, as the plots suggest. The null hypothesis is that the difference of the means is 0, and the alternative hypothesis that it is positive.

```
t.test(filter(tg, supp == "OJ", dose == 0.5)[, 1],
       filter(tg, supp == "VC", dose == 0.5)[, 1], alternative = "g")$p.value
```

```
## [1] 0.003179
```

If the needed confidence interval for the null hypothesis to be true is 95%, $\alpha = 0.05$, and since the p-value is less than α the null hypothesis is rejected. As we expected from the figures, the orange juice method is better than the ascorbic acid one with a confidence of 95% (actually a much greater confidence, since the p-value is much smaller than 0.05).

We should also test whether the orange juice is better when using 1 mg doses. With the same hypotheses than in the previous case, the p-value is

```
t.test(filter(tg, supp == "OJ", dose == 1)[, 1],
       filter(tg, supp == "VC", dose == 1)[, 1], alternative = "g")$p.value
```

```
## [1] 0.0005192
```

Again, we must reject the null hypothesis, so the orange juice is better than the ascorbic acid when using 1 mg doses. In this case the p-value is even smaller, since although the difference of the mean is similar, the variance of the OJ mean is smaller than in the 0.5 mg case, as can be seen in the `bysuppdose` data frame.

The last test between methods is the 2 mg dose case. Now, the null hypothesis is again that the difference of the mean is 0, but since we have no evidence from the previous exploratory analysis that any method is better, the alternative hypothesis will be that the difference is different from 0.

```
t.test(filter(tg, supp == "OJ", dose == 2)[, 1],
       filter(tg, supp == "VC", dose == 2)[, 1])
```

```
##
##  Welch Two Sample t-test
##
## data:  filter(tg, supp == "OJ", dose == 2)[, 1] and filter(tg, supp == "VC", dose == 2)[, 1]
## t = -0.0461, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.798  3.638
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

In this case the p-value is higher than $\alpha = 0.05$, so we cannot reject the null hypothesis. Thus, we must conclude that there is no difference between the teeth grow of guinea pigs when delivering vitamin C on orange juice or ascorbic acid. In this case I have decided to show the full `t.test()` outcome. We can see that the 95% confidence interval is -3.7981, 3.6381. This means that the probability that the interval contains the true difference in mean (unknown to us) is 95%. Since 0 is in that interval, we cannot say with a 95% confidence that the true mean is not 0. Furthermore, the interval is centered at a value close to 0, so the probability that 0 is the mean is even higher. The p-value also yields that information. It is close to 1, indicating that finding the mean difference we obtained with this data if the null hypothesis is true is very likely.

Finally, we can test whether the 2 mg dose orange juice teeth growth is significantly larger than the 1 mg one. The null hypothesis is that the difference in means is 0, and the alternative hypothesis is that the 2 mg dose mean is greater than the 1 mg mean.

```
t.test(filter(tg, supp == "OJ", dose == 2)[, 1],
       filter(tg, supp == "OJ", dose == 1)[, 1], alternative = "g")$p.value
```

```
## [1] 0.0196
```

The p-value is less than $\alpha = 0.05$, so we must reject the null hypothesis. There is significant statistical evidence that the 2 mg dose is better than the 1 mg dose.

Conclusions and assumptions

Although not stated previously, some assumptions have been made when performing the t-tests. First, I have assumed that each set of guinea pigs is different, so we should perform unpaired t-tests. I have also assumed that the variances are different. Actually, the variances have been calculated and appear in the `bysuppdose` data frame, and they are in fact different. Finally, as said in the first t-test, the confidence level is the standard 95%. The last t-test, that determines whether there is significant difference in delivering 1 or 2 mg doses with orange juice, would yield a different result if a more conservative confidence level, for example 99%, were used. Of course, the appropriate confidence level depends on how the results are going to be interpreted and used. If, for example, we would want to grow guinea pigs teeth and vitamin C in orange juice were very expensive, we could ask for a higher confidence interval, to be sure that it is worth to pay twice as much money.

As a conclusion, we have confirmed with the t-tests what we saw earlier during the exploratory analysis. The orange juice method is better for both the 0.5 and 1 mg doses, but there is no significant difference between the juice and the ascorbic acid for the 2 mg dose. Also, the t-tests have cast light on the mean doubt that the initial analysis could not answer. The orange juice 2 mg produces significantly greater growth than the orange juice 1 mg (and although I have not showed it here, the ascorbic acid 2 mg does too). However, as we have seen, a slightly more conservative confidence interval can lead us to dismiss this difference.