Statistical Inference Course Project-2

Shivangi

8/8/2020

Part 2: Basic Inferential Data Analysis Instructions

Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package.

Loading the data.

```
library(datasets)
td <- ToothGrowth
head(td)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
nrow(td)
```

```
## [1] 60
```

```
summary(td)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

From the description of this dataset (run "?ToothGrowth" in the console), we can tell that this dataset consists of 60 observations. Each observation corresponds to a guinea pig that gave its informed consent to being experimented on. The experiment consisted of giving the subjects Vitamin C supplements and examining the length of their tooth growth cells (odontoblasts) after some unspecified period of time.

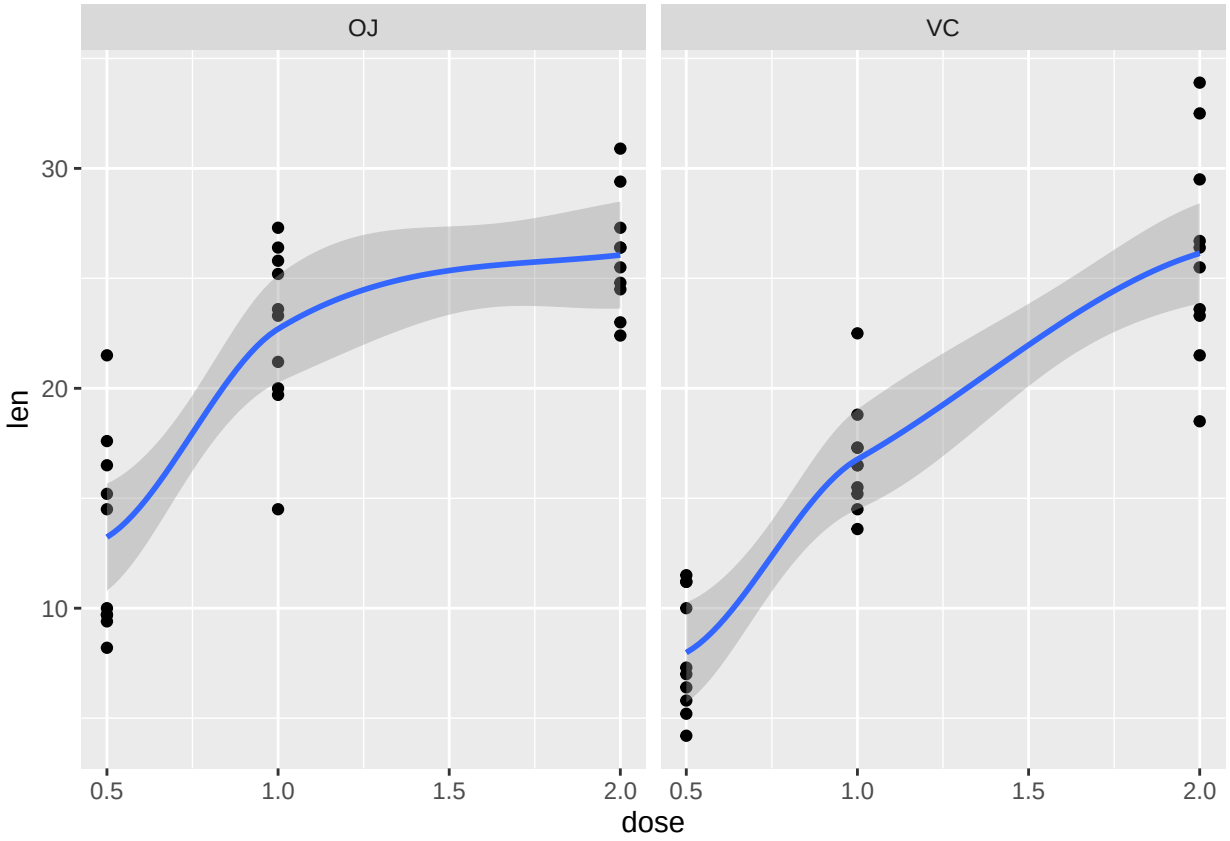
Each observation consists of three variables:

dose: The dose of vitamin C given.
supp: The method of delivering the dose. This is a factor
variable with two values: VC (ascorbic acid) and OJ (orange juice).
len: The length of the cell (not 100% sure about the units).

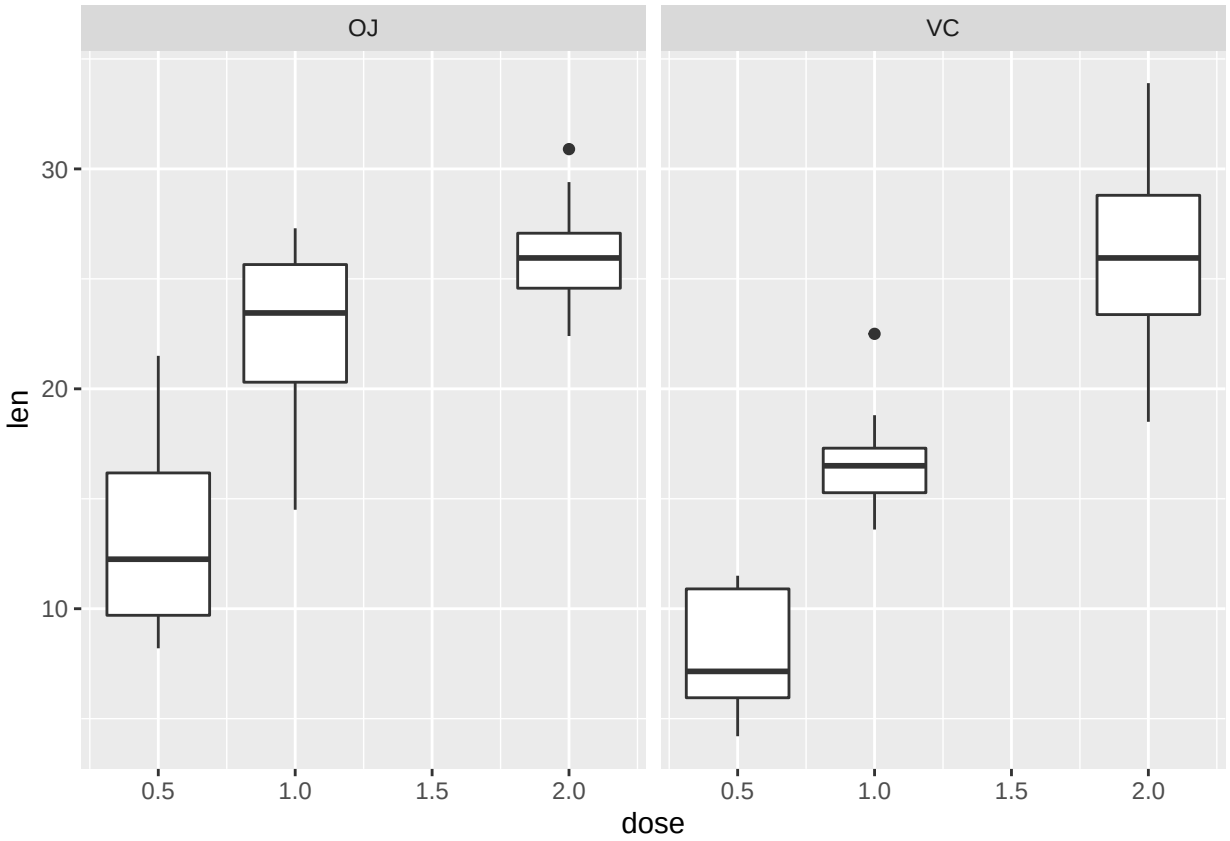
Let's visualize this dataset:

```
library(ggplot2)
ggplot(data = td) +
    geom_point(mapping=aes(x=dose, y=len)) +
    facet_wrap(~ supp, nrow=1) +
    geom_smooth(mapping=aes(x=dose, y=len))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
ggplot(data = td, aes(dose, len)) +
  geom_boxplot(mapping=aes(group=dose)) +
  facet_wrap(~ supp, nrow=1)
```



Comarisons and Confidence Intervals

We want to compare the two vitamin C delivery methods (OJ and VC) in terms of their effect on the cell length. Assuming tooth growth is a good thing, we want to maximize the effect of delivering a particular dose. The question we want to ask first is: which is the more effective delivery method? Since there are three dosages, we effectively have three datasets to compare.

Let's start with the lowest dosage: 0.5mg/day. We will set up a hypothesis test under the following conditions:

Null: both delivery methods are equally efficient Alternative: there is a measurable difference between the two methods Since the sample size is relatively small (10 samples per delivery method), we will apply a two-sided t-test. The observations refer to different subjects, so we must use an unpaired test.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
oj_low = filter(td, supp == 'OJ', dose == .5)$len
vc_low = filter(td, supp == 'VC', dose == .5)$len
t.test(oj_low - vc_low, alternative='two.sided',
       paired=FALSE, conf.level = .95)
```

```
##
##  One Sample t-test
##
## data:  oj_low - vc_low
## t = 2.9791, df = 9, p-value = 0.01547
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.263458 9.236542
## sample estimates:
## mean of x
##      5.25
```

The results show us that the mean difference between the lengths is 5.5 units in favor of the OJ method. The 95% confidence interval ranges from 1.2 to 9.2 units, allowing us to reject the null hypothesis. For smaller dosages, OJ is the clear winner.

Let's repeat the above for the middle dosage (1.0mg/day):

```r
oj_mid = filter(td, supp == 'OJ', dose == 1.)$len
vc_mid = filter(td, supp == 'VC', dose == 1.)$len
t.test(oj_mid - vc_mid, alternative='two.sided',
       paired=FALSE, conf.level = .95)
```

```
##
##  One Sample t-test
##
## data:  oj_mid - vc_mid
## t = 3.3721, df = 9, p-value = 0.008229
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.951911 9.908089
## sample estimates:
## mean of x
##      5.93
```

Again, OJ is the clear winner here: the CI is 1.95 to 9.9, and we reject the null.

Finally, let's examine the high dosage (2.0mg/day):

```r
oj_hi = filter(td, supp == 'OJ', dose == 2.)$len
vc_hi = filter(td, supp == 'VC', dose == 2.)$len
t.test(oj_hi - vc_hi, alternative='two.sided', paired=FALSE, conf.level = .95)
```

```
##
##  One Sample t-test
##
## data:  oj_hi - vc_hi
```

```
## t = -0.042592, df = 9, p-value = 0.967
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -4.328976  4.168976
## sample estimates:
## mean of x
##     -0.08
```

In this case, the mean difference is close to zero. The CI also includes zero, meaning we fail to reject the null hypothesis. This means there is no clear winner in this particular case.

```
c(var(oj_hi), var(vc_hi))
```

```
## [1]  7.049333 23.018222
```

Another important observation is the variance in results for the high dose, which is evident from eyeballing the plot. More precisely, shows us the VC method indeed yields results of greater variance (more than three times greater).

Conclusions:

The OJ delivery method was more effective for low and mid dosages. For the high dosage, both methods performed approximately the same. However, from looking at the scatter plot above, it is evident that the OJ plot is plateauing, while the VC plot is continuing to increase. This suggests that if the experiment included dosages higher than 2.0mg/day, the VC delivery method would be more effective.

However, the VC delivery method also yielded results of greater variance at higher dosages. This may be a relevant factor to consider when choosing between the two methods.

This conclusion depends on the assumption that the samples were obtained from independent populations.