Statistical Inference Course Project - Part 2

John Letteboer

10/21/2018

Synopsis

This is the project for the statistical inference class. In it, you will use simulation to explore inference and do some simple inferential data analysis. The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis.

We're going to analyze the ToothGrowth data in the R datasets package. You should

1. Load the ToothGrowth data and perform some basic exploratory data analyses
2. Provide a basic summary of the data.
3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)
4. State your conclusions and the assumptions needed for your conclusions.

Question 1

Load the ToothGrowth data and perform some basic exploratory data analyses

```
data(ToothGrowth)
tg <- ToothGrowth
# convert dose to factor
tg$dose <- as.factor(tg$dose)

# string
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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

```
# head
head(tg,2)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
```

```
# table
table(tg$dose, tg$supp)
```

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

Question 2

Provide a basic summary of the data.

```
# summary
summary(tg)
```

```
##       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
```

Let's plot the data

```
library(ggplot2)
require(gridExtra)
```

```
## Loading required package: gridExtra
```

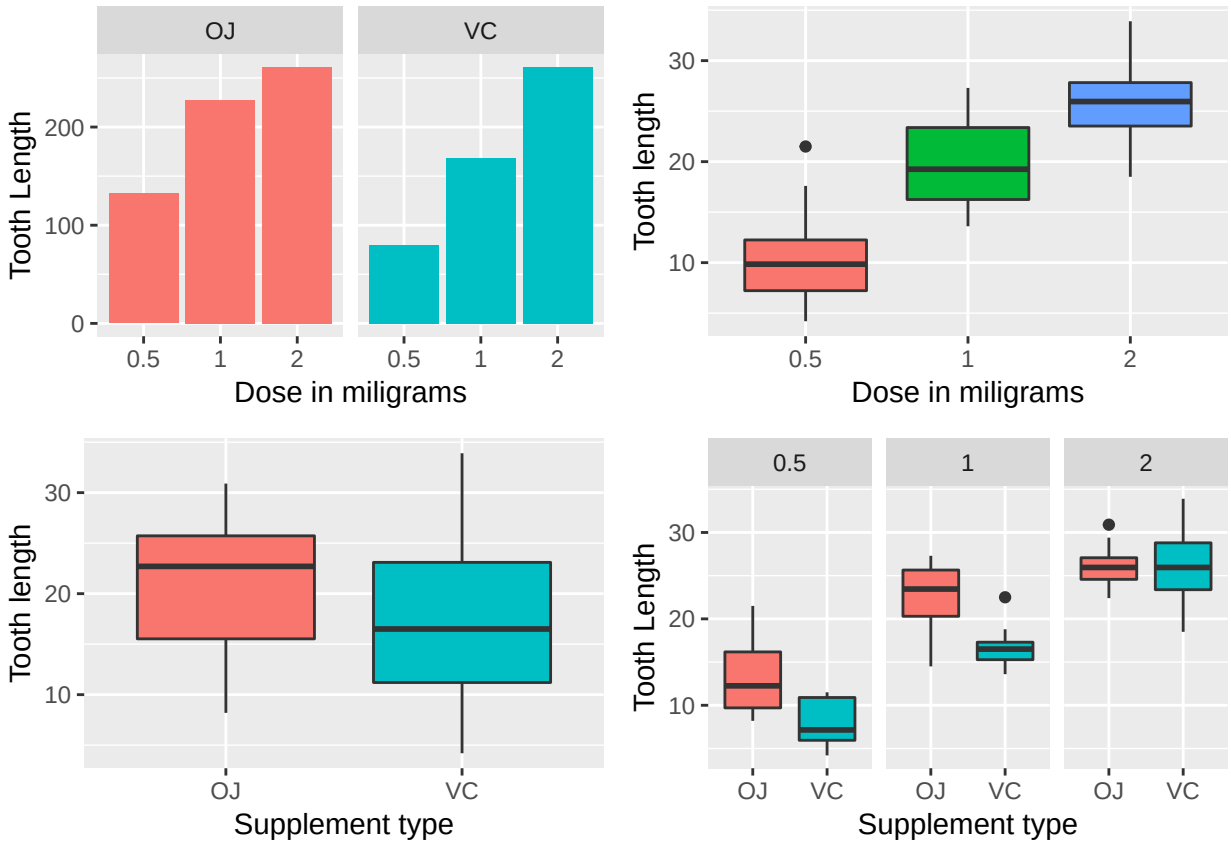
```
plot1 <- ggplot(data=tg, aes(x=dose, y=len, fill=supp)) +
    geom_bar(stat="identity") +
    facet_grid(. ~ supp) +
    xlab("Dose in miligrams") + ylab("Tooth Length") +
    guides(fill=FALSE)

plot2 <- ggplot(data=tg, aes(x=dose, y=len, fill=dose)) +
    geom_boxplot() +
    xlab("Dose in miligrams") + ylab("Tooth length") +
    guides(fill=FALSE)

plot3 <- ggplot(data=tg, aes(x=supp, y=len, fill=supp)) +
    geom_boxplot() +
    xlab("Supplement type") + ylab("Tooth length") +
    guides(fill=FALSE)

plot4 <- ggplot(data=tg, aes(x=supp, y=len, fill=supp)) +
    geom_boxplot() +
    facet_grid(. ~ dose) +
    xlab("Supplement type") + ylab("Tooth Length") +
    guides(fill=FALSE)

grid.arrange(plot1, plot2, plot3, plot4, ncol=2)
```



Question 3

Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)
What statustical test are we going to use, we could use T-test or Analysis of Variance (ANOVA). T-test is limited to comparing means of two groups, one-way ANOVA can compare more than two groups.

Let's do some test based on Analysis of Variance (ANOVA).

```
an <- aov(len ~ supp * dose, data=tg)
summary(an)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## supp         1  205.4   205.4  15.572 0.000231 ***
## dose         2 2426.4  1213.2  92.000  < 2e-16 ***
## supp:dose    2  108.3    54.2   4.107 0.021860 *
## Residuals   54  712.1    13.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This table confirms that there are differences between the groups. We see the following between the groups:

`len` and `supp` = `F(1,54)=15.572;p<0`
`len` and `dose` = `F(2,54)=92.000;p<0`
`supp` and `dose` = `F(2,54)=4.107;p<0.05`

There is a minor interaction between the combination of supplement type `supp` and dosage `dose`.

Let's do a Post Hoc test with Tukey HSD (Honestly Significant Difference).

```
TukeyHSD(an)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = len ~ supp * dose, data = tg)
##
## $supp
##       diff       lwr       upr     p adj
## VC-OJ -3.7 -5.579828 -1.820172 0.0002312
##
## $dose
##         diff       lwr       upr   p adj
## 1-0.5  9.130  6.362488 11.897512 0.0e+00
## 2-0.5 15.495 12.727488 18.262512 0.0e+00
## 2-1    6.365  3.597488  9.132512 2.7e-06
##
## $`supp:dose`
##                 diff        lwr        upr     p adj
## VC:0.5-OJ:0.5 -5.25 -10.048124 -0.4518762 0.0242521
## OJ:1-OJ:0.5    9.47   4.671876 14.2681238 0.0000046
## VC:1-OJ:0.5    3.54  -1.258124  8.3381238 0.2640208
## OJ:2-OJ:0.5   12.83   8.031876 17.6281238 0.0000000
## VC:2-OJ:0.5   12.91   8.111876 17.7081238 0.0000000
## OJ:1-VC:0.5   14.72   9.921876 19.5181238 0.0000000
## VC:1-VC:0.5    8.79   3.991876 13.5881238 0.0000210
## OJ:2-VC:0.5   18.08  13.281876 22.8781238 0.0000000
## VC:2-VC:0.5   18.16  13.361876 22.9581238 0.0000000
## VC:1-OJ:1     -5.93 -10.728124 -1.1318762 0.0073930
## OJ:2-OJ:1      3.36  -1.438124  8.1581238 0.3187361
## VC:2-OJ:1      3.44  -1.358124  8.2381238 0.2936430
## OJ:2-VC:1      9.29   4.491876 14.0881238 0.0000069
## VC:2-VC:1      9.37   4.571876 14.1681238 0.0000058
## VC:2-OJ:2      0.08  -4.718124  4.8781238 1.0000000
```

There are significant differences between each of the groups in `supp:dose`. These are not significant:

`VC:0.5-OJ:0.5`, `VC:1-OJ:0.5`, `OJ:2-OJ:1`, `VC:2-OJ:1` and `VC:2-OJ:2`

The function confint is used to calculate confidence intervals on the treatment parameters, by default 95% confidence intervals:

```
confint(an)
```

```
##                    2.5 %    97.5 %
## (Intercept)  10.9276907 15.532309
## suppVC       -8.5059571 -1.994043
## dose1         6.2140429 12.725957
## dose2         9.5740429 16.085957
## suppVC:dose1 -5.2846186  3.924619
## suppVC:dose2  0.7253814  9.934619
```

And computes the summary tables for model fits of the mean response for each combinations of levels of the factors in a term

```
print(model.tables(an,"means"),digits=3)
```

```
## Tables of means
## Grand mean
##
## 18.81333
##
##  supp
## supp
##    OJ    VC
## 20.66 16.96
##
##  dose
## dose
##   0.5     1     2
## 10.60 19.73 26.10
##
##  supp:dose
##     dose
## supp 0.5   1     2
##   OJ 13.23 22.70 26.06
##   VC  7.98 16.77 26.14
```

Question 4

State your conclusions and the assumptions needed for your conclusions.

Conclusions

1. Length and supplement, the p-value is less than 0.05, so supplement alone type has effect on tooth growth.
2. Length and dosage, the p-value is less the 0.05, so dosage alone has effect on tooth growth.
3. Lendth and a combination of supplement and dosage, the p-value is also less than 0.05, there is a minor interaction.

Supplement type has a influence, but `OJ` has a greater average teethgrowth in combination with dosages `0.5` and `1` then for the `VC` supplement, while teeth length for the `VC` supplement and `OJ` in combiantion with dosage `2` has no significant effect (almost the same mean & confidence interval).

Assumptions

- The experiment was done with random assignment of guinea pigs to different dose level categories and supplement type to control for confounders that might affect the outcome.
- Members of the sample population, i.e. the 60 guinea pigs, are representative of the entire population of guinea pigs. This assumption allows us to generalize the results.
- The data is normal distributed.