Statistical Inference on ToothGrowth Dataset

Gordon CHAN

19/08/2015

Overview

This report will demonstrate the Central Limit Theorem (CLT) through a basic inferential data analysis.

Dataset

The default dataset from R, ToothGrowth is loaded.

```
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
```

```
# library("reshape2")
library("ggplot2")
    library("gridExtra")
    library("ggthemes")
library("knitr")

data(ToothGrowth)
    head(ToothGrowth)
```

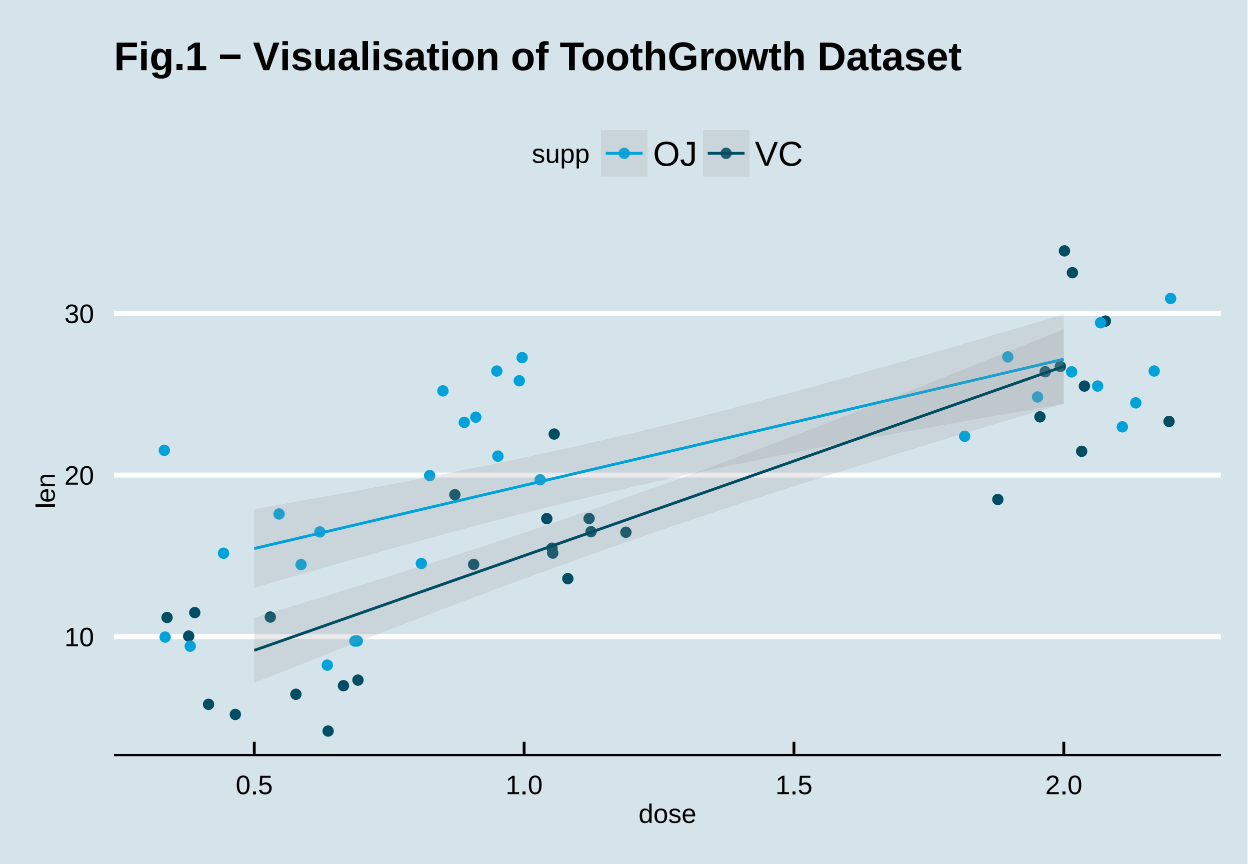
```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

From the documentation available from the R package, the ToothGrowth dataset is titled as *The Effect of Vitamin C on Tooth Growth in Guinea Pigs*. Each entry correspond to a guinea pig, each received one of three dose level of Vitamin C, via one of two delivery methods. *len* is a numeric variable of the length of odonthoblasts. *supp* is a factor variable of the supplement delivery type, either as **VC** (Ascorbic Acid) or **OJ** (Orange Juice). *dose* is a numeric variable of the dosage of Vitamin C in milligrams, at **0.5, 1, or 2mg**.

Exploratory Data Analysis

The raw dataset is plotted to visualise the data, regression lines are also added to the plot. The data points are plotted as jitters rather than points to avoid overplotting.

```
s <- ggplot(ToothGrowth, aes(x=dose, y=len, colour=supp))+
    geom_jitter()+
    geom_smooth(method="lm", alpha=0.2)+
        labs(title="Fig.1 - Visualisation of ToothGrowth Dataset")+
            theme_economist()+
            scale_color_economist()
s
```

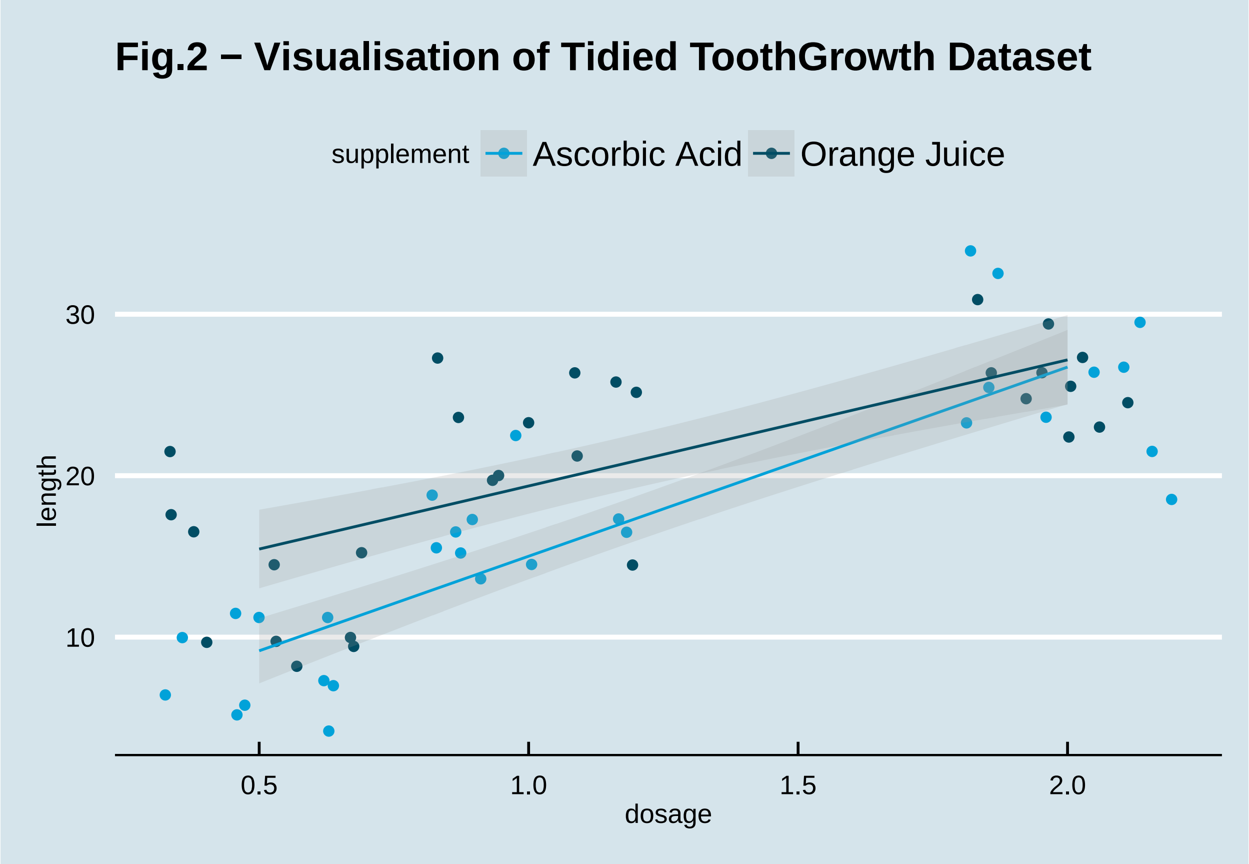


The name of the variables are tidied up, by replacing with meaningful names.

```
names(ToothGrowth) <- c("length", "supplement", "dosage")

ToothGrowth$supplement <- sapply(ToothGrowth$supplement, gsub,
                                 pattern="OJ", replacement="Orange Juice")
ToothGrowth$supplement <- sapply(ToothGrowth$supplement, gsub,
                                 pattern="VC", replacement="Ascorbic Acid")

s <- ggplot(ToothGrowth, aes(x=dosage, y=length, colour=supplement))+
    geom_jitter()+
    geom_smooth(method="lm", alpha=0.2)+
        labs(title="Fig.2 - Visualisation of Tidied ToothGrowth Dataset")+
            theme_economist()+
            scale_color_economist()
s
```



A summary table is calcualted. The average legnth of odontoblasts by different dosage of Vitamin C and by supplement type is shown. We can see that for each dosage group, the average length for the guinea pigs supplemented through orange juice appeared generally higher than those supplemented through ascorbic acid.

```
TG_summary <- ToothGrowth %>% select(supplement, dosage, length) %>%
    group_by(supplement, dosage) %>% summarise(average_length = mean(length)) %>%
    ungroup() %>% select(dosage, supplement, average_length) %>% arrange(dosage, supplement)

kable(TG_summary, caption="Summary Table of the ToothGrowth dataset")
```

Table 1: Summary Table of the ToothGrowth dataset

dosage	supplement	average_length
0.5	Ascorbic Acid	7.98
0.5	Orange Juice	13.23
1.0	Ascorbic Acid	16.77
1.0	Orange Juice	22.70
2.0	Ascorbic Acid	26.14
2.0	Orange Juice	26.06

Statistical Inference

To test whether the apparent differences in odonthoblasts length are statistically significant, t-tests are performed between the legnth and supplement for each dosage group. For the null hypothesis (H0), the true difference in means of odonthoblasts length, between different supplement type is equal to 0. While for the alternative hypothesis (H1), the true difference in means of odonthoblasts length, between different supplement type is **NOT** equal to 0. The hypothesis testing is repeated for each of the dosage group (0.5mg, 1mg, and 2mg).

```
TG0.5 <- subset(ToothGrowth, dosage==0.5)

    tt0.5 <- t.test(length ~ supplement, paired = FALSE, var.equal = FALSE, data = TG0.5)
        tt0.5
```

```
##
##  Welch Two Sample t-test
##
## data:  length by supplement
## t = -3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.780943 -1.719057
## sample estimates:
## mean in group Ascorbic Acid  mean in group Orange Juice
##                        7.98                       13.23
```

```
TG1 <- subset(ToothGrowth, dosage==1)

    tt1 <- t.test(length ~ supplement, paired = FALSE, var.equal = FALSE, data = TG1)
        tt1
```

```
##
##  Welch Two Sample t-test
##
## data:  length by supplement
## t = -4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.057852 -2.802148
## sample estimates:
## mean in group Ascorbic Acid  mean in group Orange Juice
##                       16.77                       22.70
```

```
TG2 <- subset(ToothGrowth, dosage==2)

    tt2 <- t.test(length ~ supplement, paired = FALSE, var.equal = FALSE, data = TG2)
        tt2
```

```
##
##  Welch Two Sample t-test
##
## data:  length by supplement
## t = 0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.63807  3.79807
## sample estimates:
## mean in group Ascorbic Acid  mean in group Orange Juice
##                       26.14                       26.06
```

Conclusion

The results from the t-tests are summarised as a table. The significance for each p-value are also generated, where if the p-value is less then 1% the difference is **highly significant**; if the p-value is between 1% to 5% the difference is **significant**; if the p-value is larger or equal to 5% the difference is **not significant**.

For the dosage group at 0.5mg of Vitamin C, with a p-value of just 0.64%, there are highly significant difference between the mean in odonthoblasts length between those treated with Ascorbic Acid and Orange Juice. Hence the null hypothesis is rejected.

For the dosage group at 1mg of Vitamin C, with a p-value of just 0.1%, there are highly significant difference between the mean in odonthoblasts length between those treated with Ascorbic Acid and Orange Juice. Hence the null hypothesis is rejected.

However, for the dosage group at 2mg of Vitamin C, with a p-value of 96%, there are **NO** significant difference between the mean in odonthoblasts length between those treated with Ascorbic Acid and Orange Juice. Hence the null hypothesis failed to be rejected.

```r
result <- data.frame(cbind(
    distinct(TG_summary, dosage)$dosage,
    c(tt0.5$estimate[[1]]-tt0.5$estimate[[2]],
      tt1$estimate[[1]]-tt1$estimate[[2]],
      tt2$estimate[[1]]-tt2$estimate[[2]]),
    c(tt0.5$p.value, tt1$p.value, tt2$p.value)
))
        names(result) <- c("dosage", "difference(VC-OJ)", "p_value")

sig <- function(x){
    ifelse(x<0.01, "highly significant", ifelse(x<0.05, "significant", "not significant"))
}

        result$significance <- sapply(result$p_value, sig)

    kable(result, caption="Inference Result for Each Dosage Group")
```

Table 2: Inference Result for Each Dosage Group

dosage	difference(VC-OJ)	p_value	significance
0.5	-5.25	0.0063586	highly significant
1.0	-5.93	0.0010384	highly significant
2.0	0.08	0.9638516	not significant

Assumptions

We assumed the data are independent and identically distributed (iid) for the purpose of applying the t-test. In addition, we assumed that the variance of the data are not equal.