Coursera Statistical Inference Course Project Part 2

kimnewzealand

29 May 2017

Overview

In part 2 of this this project we are performing basic inferential data analysis using the ToothGrowth data in the R datasets package.

For this analysis, the following R package is needed `ggplot2`.

```
library(ggplot2)
```

Part 2: Basic Inferential Data Analysis Instructions

1. Load the ToothGrowth data

```
# Load the ToothGrowth data
data(ToothGrowth)
```

2. Provide a basic summary of the data and perform some basic exploratory data analysis

```
# Look at the structure of 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 ...
```

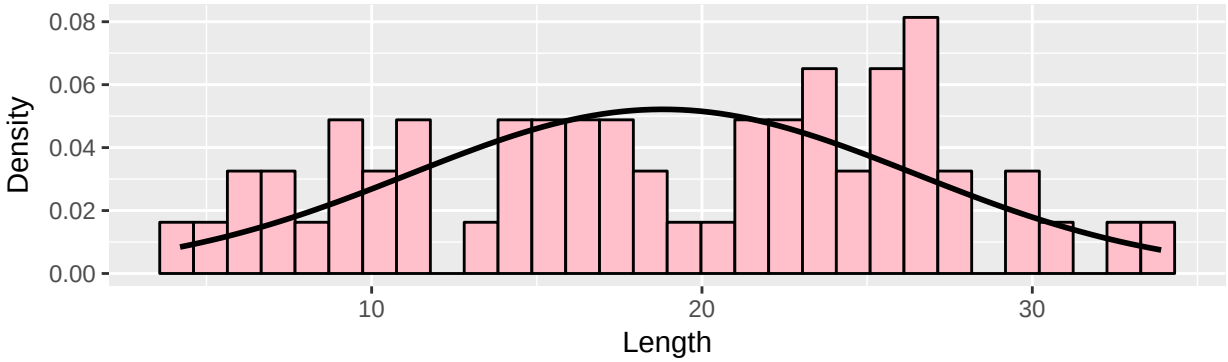
```
# Calculate summary statistics of the len variable, including sample mean
summary(ToothGrowth$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20   13.08   19.25   18.81   25.28   33.90
```

Referring to the R documentation, the response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. 10 guinea pigs were assigned to each of three dosages. Each group received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) delivered by two delivery methods, (orange juice or ascorbic acid (a form of vitamin C and coded as VC).

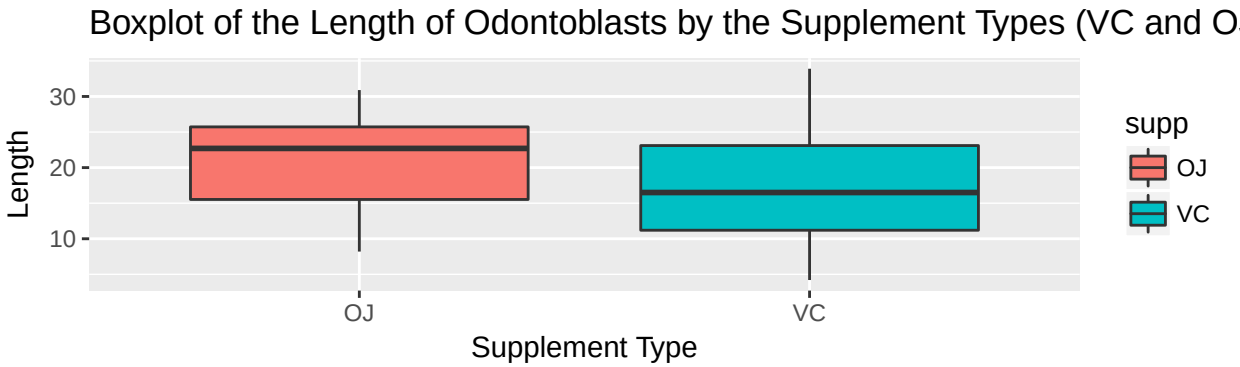
```
# Plot histogram of the numerical len variable to view the distribution
g <- ggplot(ToothGrowth,aes(len))
g + geom_histogram(aes(y = ..density..),colour="black",fill="pink")+
       stat_function(fun=dnorm,args=list( mean=mean(ToothGrowth$len), sd=sqrt(var(ToothGrowth$len))),ge
      scale_x_continuous("Length")+
      ylab("Density")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



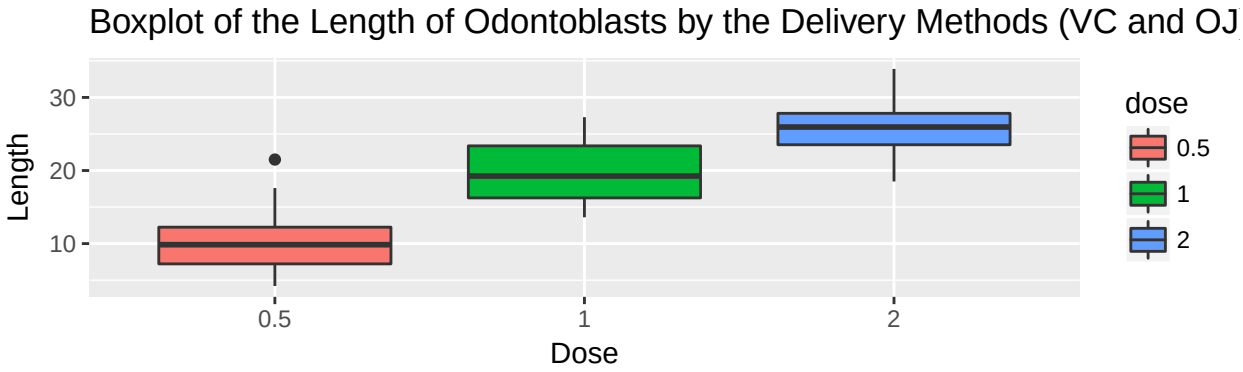
The shape of the histogram bars is not easily seen to be the same as the calculated normal distribution in black. This may be explained by the grouping of the guinea pigs receiving different dose levels. We will explore the len versus the other two variables further.

```
# View the relationship between the numerical len and the categorical supp variables
# using a boxplot
ToothGrowth$supp <- as.character(ToothGrowth$supp)
g <- ggplot(ToothGrowth,aes(supp,len))
g+geom_boxplot(aes(fill=supp))+
      ggtitle("Boxplot of the Length of Odontoblasts by the Supplement Types (VC and OJ)")+
      xlab("Supplement Type")+
      ylab("Length")
```



The OJ median is slightly higher than the VC median in the boxplot plotted against the len variable. The OJ distribution appears slightly left skewed as the median is closer to the 75% percentile, but the VC looks normal with a centred median. Both the OJ and VC supp have similar variability.

```
# View the relationship between the numerical variable lenand categorical variable the dose using a box
ToothGrowth$dose <- as.character(ToothGrowth$dose)
g <- ggplot(ToothGrowth,aes(dose,len))
g + geom_boxplot(aes(fill=dose))+
      ggtitle("Boxplot of the Length of Odontoblasts by the Delivery Methods (VC and OJ)")+
      xlab("Dose")+
      ylab("Length")
```



The median of the doses increases as the len increases in the boxplot plotted against the len variable. The distributions are approximately normal as the medians are centred, but dose have similar variability.

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)

First set the hypothesis test:

The null hypothesis $H_0 : \mu = 0$ is that population mean difference (μ) is not different of the length by the supp, given a dose level.

The alternative hypothesis $H_1 : \mu \neq 0$ is that there is a population mean difference of length by supp, given a dose level.

Next check the conditions of independence and skewness of the variables in the dataset based on the sample statistics and the distributions of the variables.

Since the population standard deviation is unknown we can look to use the t-distribution for inference of a population mean and confidence interval and calculate statistical significance.

The t-distribution assummes that the data are independent and identically distributed (iid).

As we have 60 observations from 60 different guinea pigs, we cannot use the paired group test. We can use the t.test function in R to test differences in means from independent groups assuming equal variance.

```
# Perform a t test using t.test() function for each dose level and print out p-Values
dose_group<-levels(factor(ToothGrowth$dose))
reject<-paste(" ")
notreject<-paste(" ")
for (level in dose_group){
    result<-t.test(len ~ supp, ToothGrowth[ToothGrowth$dose == level, ])
    print(paste("For dose",as.character(level)," the t.test result is: "))
    print(result)
    ifelse(result$p.value<0.05,
           print(paste("Since p-value < 5%, reject null hypothesis for dose level",as.character(level))),
           print(paste("Since p-value > 5%, fail to reject null hypothesis for dose level ",as.characte
    ifelse(result$p.value<0.05,
           reject<-paste(reject," and ",as.character(level)),
           notreject <- paste(notreject," and  ",as.character(level)))
    "\n"
    }
```

[1] "For dose 0.5 the t.test result is:"

```
Welch Two Sample t-test
```

data: len by supp 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: 1.719057 8.780943 sample estimates: mean in group

OJ mean in group VC 13.23 7.98

[1] “Since p-value < 5%, reject null hypothesis for dose level 0.5” [1] “For dose 1 the t.test result is:”

```
Welch Two Sample t-test
```

data: len by supp 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: 2.802148 9.057852 sample estimates: mean in group OJ mean in group VC 22.70 16.77

[1] “Since p-value < 5%, reject null hypothesis for dose level 1” [1] “For dose 2 the t.test result is:”

```
Welch Two Sample t-test
```

data: len by supp 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: -3.79807 3.63807 sample estimates: mean in group OJ mean in group VC 26.06 26.14

[1] “Since p-value > 5%, fail to reject null hypothesis for dose level 2”

4. State your conclusions and the assumptions needed for your conclusions.

We conclude form the analysis the following:

- In the initial exploratory data analysis, the supplment type (supp) was not likely a significant factor however the doses was possibly a factor in the length of odontoblasts in Guinea Pigs.
- The t-test results show a 95% confidence interval which can be interpreted that we are 95% confident that the true mean difference in length of odontoblasts (len) by orange juice (OJ) or ascorbic acid (VC) supplement methods (supp) given the three dose levels (dose) are in this interval. It also provides the critical value,t, the degrees of freedom,df and p-value and sample mean of the differences.
- The p-value, which the probability of observing an outcome as extreme probability of finding the observed, or more extreme, results when the null hypothesis (H_0) of a study question is true. As a rule of thumb, where the p-value is less than 5%, we have strong evidence against the null hypothesis.
- We reject the null hypothesis H_0 for difference in supp, VC and OJ given dose and 0.5 and 1. However we do not reject the difference given dose level and 2.
- This conclusion assumes conditions of a t-test that the underlying data are independent and are normally distributed.