DSDM2015

Coursera Data Science Specialization, Statistical Inference

Course Project Part 2: Basic Inferential Data Analysis

Introduction

The data set consists of 60 observations for the tooth growth in guinea pigs. The guinea pigs received two different treatments and three different levels of dosage. In particular, the guinea pigs received Vitamin C in two different forms (orange juice, ascorbic acid) and at the levels of dosage (0.5, 1, and 2 mg). There were 10 guinea pigs in each group. The question is whether tooth growth differed between the groups receiving different treatments in terms of the delivery or dosage.

Data Summary and Exploratory Data Analysis

The length of the tooth ranged from 4.2 to 33.9 with a mean of 18.81 and a standard deviation of 7.65. Figure 1 shows a histogram of tooth length with a fitted normal distribution (blue line). The distribution is roughly symmetrical, but seems to be multi-modal. This is a first hint that there might be different groups within the data. There are no indications of outliers. In fact, the mean values of tooth length do differ across the different levels of dosage and across the different type of treatment

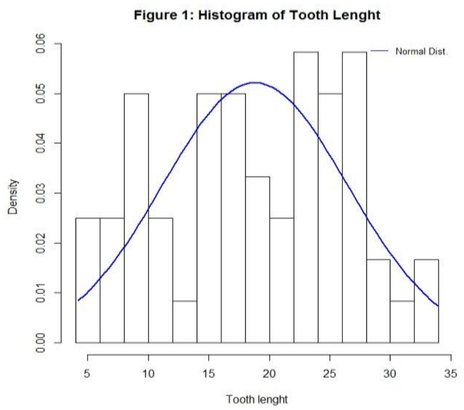


Table 1

Group	*Mean of tooth length*
Dosage	
0.5	10.61
1	19.74
2	26.1
Delivery	
Orange Juice (OJ)	20.66
Ascorbic acid (VC)	16.96

The mean tooth length is higher for orange juice than for ascorbic acid. A higher dosage is also associated with a larger mean value for tooth length (Table 1).

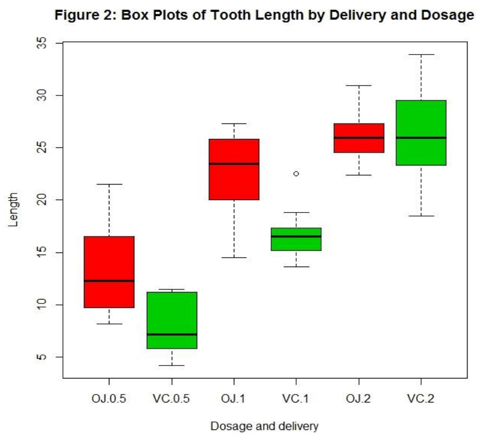


Figure 2 shows box plots for all the different combinations of dosage and delivery. Again, orange juice (OJ) is associated with larger teeth then ascorbic acid (VC), however with a dosage of 2 this difference is quite small. Higher dosages are also associated with larger teeth for both types of delivery. There are no indications of outliers (except for possible outlier for ascorbic acid at dosage 1). The variances differ across the groups.

__Estimating the Effect of the Treatment: Confidence intervals and Hypotheses Tests__

The exploratory data analysis points to differences of teeth growth among the different treatments in terms of delivery (orange juice, ascorbic acid) and dosage. In particular, a higher dosage and orange juice as a form of delivery are associated with larger teeth. This difference between samples might be due to random sampling variation, however. We can calculate confidence intervals and perform hypothesis checks to investigate this possibility.

The data comes from a (presumably) random assignment of treatments to different guinea pigs, thus there are several independent groups. The data seems to be roughly symmetrical with no outliers, which makes a t-test for the difference means suitable. Figure 2 indicates that the variances of the groups differ, thus we cannot assume equal variance. We have no reason to a priori rule out that one group has in fact a higher mean than the other, thus a two-sided test is appropriate. We will use the customary 5% threshold to assess the statistical significance. The null hypothesis is that there is no difference in the population mean values between the various groups.

Table 2: Confidence Intervals and P-Values for Tests of Mean Difference (two-sided, alpha=0.05)			
Treatment groups	t-Value	p-Value	95%-Confidence interval
Delivery (OJ vs. VC)	1.915	0.06	-0.171, 7.571
Delivery (OJ vs. VC) (Dosage=0.5)	3.17	0.006	1.719, 8.78
Delivery (OJ vs. VC) (Dosage=1)	4.033	0.001	2.802, 9.058
Delivery (OJ vs. VC) (Dosage=2)	-0.46	0.964	-3.798, 3.638
Dosage (0.5 vs. 1)	-6.477	<0.001	-11.984, -6.276
Dosage (1 vs. 2)	-4.9	<0.001	-8.996, -3.734
Dosage (0.5 vs. 1)	-11.799	<0.001	-18.156, -12.834

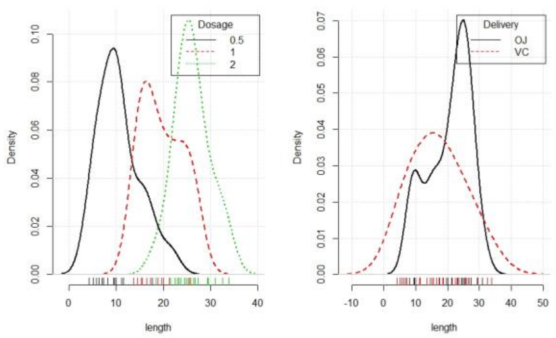
Table 2 reports the t-values, p-values and 95% confidence intervals for the differences of means between the various treatment groups. The comparison across type of delivery for all dosages yields a p-value of 0.06 and a confidence interval that includes zero. Thus, we cannot rule out (at the 5% level) that the differences in sample means are induced by chance. Figure 2 showed that the difference between delivery types was less pronounced for the highest dosage. Indeed, the p-value for the t-test is very high (0.964). Hence, we cannot rule out that the difference in sample means is due to chance variation. In contrast, at lower dosages (0.5, 1) there is a clear difference between orange juice and ascorbic acid. The p-values are well below our 5% threshold and the confidence intervals for the mean difference are not close to zero. There is also a statistically significant difference of tooth length between the different treatments in terms of dosage. The p-values are very far from our 5% threshold and the confidence intervals are far from zero.

In sum, the results of the confidence intervals and hypotheses tests are that we can reject the null hypothesis of no difference between the groups for the different dosages. In other words, the differences between tooth length between different treatments in terms of dosage are statistically significant at the 5%-level. Regarding the delivery type, we can only reject the null hypothesis of no difference for the lower dosages (alpha=0.05), but not for all cases and for group receiving the highest dosage.

Checking assumptions

The sampling distribution of a mean follows a t-distribution (with n-1 degrees of freedom) asymptotically if the data is independent and identically distributed with a normal distribution. It still approximately holds if the data is symmetrical and does not have large outliers.

Figure 3: Kernel density estimates



As already discussed above, the overall distribution is roughly symmetrical and there do not seem to be many/large outliers. Figure 3 confirms this for the different groups. The data is roughly symmetrical. We can also assume that the data is independent due to the research design (which apparently was based on random assignment).

Thus, while the assumptions for the use of t-tests are not strictly met, the use of t-tests and t-confidence intervals is still justified.

Conclusion

Teeth length of guinea pigs is affected by different dosages and at lower levels of dosage by the type of delivery (orange juice, ascorbic acid). The exploratory data analysis already revealed difference between teeth length for the various groups, with a very small difference for the delivery type at the highest level (dosage of 2). This is confirmed by t-tests for the difference of means and related t-confidence intervals. The assumptions for the use of t-tests (independent and identical distribution, normality) are not strictly met. The distributions for the groups are independent but not normally distributed. However, they are roughly symmetrical, thus the results are properly still valid.

Appendix: R Code

```
### Statistical Inference, Assignment Part 2
#
## Exploratory data analysis
# Data summary
data("ToothGrowth")
?ToothGrowth
str(ToothGrowth)
summary(ToothGrowth)
sd(ToothGrowth$len)
aggregate(ToothGrowth$len~ToothGrowth$supp, FUN=mean)
aggregate(ToothGrowth$len~ToothGrowth$dose, FUN=mean)

# Histogram for tooth lenght
h1<-hist(ToothGrowth$len, main="Figure 1: Histogram of Tooth Lenght", xlab="Tooth lenght", breaks=15, prob=T)
x.fit<-seq(min(ToothGrowth$len), max(ToothGrowth$len), length=60)
y.fit<-dnorm(x.fit, mean=mean(ToothGrowth$len), sd=sd(ToothGrowth$len))
y.fit<-y.fit
lines(x.fit, y.fit, col="blue", lwd=2)
legend("topright", legend="Normal Dist.", lty=1, col="blue", bty="n", cex=.8)

# Boxplot comparing different group distributions
boxplot(ToothGrowth$len~interaction(ToothGrowth$supp, ToothGrowth$dose), data=warpbreaks, col=2:3, main="Figure 2: Box Plots of Tooth Length by Delivery and Dosage", xlab="Dosage and delivery", ylab="Length")

## Confidence intervals and hypothesis testing
# Delivery type
t.test(ToothGrowth$len ~ ToothGrowth$supp, paired = FALSE, var.equal = FALSE)
OJ.Half = ToothGrowth$len[ToothGrowth$supp == 'OJ' & ToothGrowth$dose == 0.5]
VC.Half = ToothGrowth$len[ToothGrowth$supp == 'VC' & ToothGrowth$dose == 0.5]
t.test(OJ.Half, VC.Half, paired = FALSE, var.equal = FALSE)
OJ.1 = ToothGrowth$len[ToothGrowth$supp == 'OJ' & ToothGrowth$dose == 1]
VC.1 = ToothGrowth$len[ToothGrowth$supp == 'VC' & ToothGrowth$dose == 1]
t.test(OJ.1, VC.1, paired = FALSE, var.equal = FALSE)
OJ.2 = ToothGrowth$len[ToothGrowth$supp == 'OJ' & ToothGrowth$dose == 2]
VC.2 = ToothGrowth$len[ToothGrowth$supp == 'VC' & ToothGrowth$dose == 2]
t.test(OJ.2, VC.2, paired = FALSE, var.equal = FALSE)
# Dosage
Dose.Half = ToothGrowth$len[ToothGrowth$dose == 0.5]
Dose.One = ToothGrowth$len[ToothGrowth$dose == 1]
Dose.Two = ToothGrowth$len[ToothGrowth$dose == 2]
t.test(Dose.Half,  Dose.One, paired = FALSE, var.equal = FALSE)
t.test(Dose.One,  Dose.Two, paired = FALSE, var.equal = FALSE)
t.test(Dose.Half,  Dose.Two, paired = FALSE, var.equal = FALSE)

## Checking normality assumption
dev.new()
par(mfrow=c(1,2))
par("oma")
par("oma"=c(0,0,2,0))
```

```
ToothGrowth$Dosage<-as.factor(ToothGrowth$dose)
ToothGrowth$Delivery<-ToothGrowth$supp
densityPlot(len ~ Dosage, data=ToothGrowth, xlab="length", bty="n")
densityPlot(len ~ Delivery, data=ToothGrowth, xlab="length", bty="n")
mtext("Figure 3: Kernel density estimates",side=3, outer=T, line=-1, cex=1.5)
```