Statistical Inference Course Project - Part 2

Overview

R includes the ToothGrowth dataset, which shows the length of the teeth of each of 10 guinea pigs after dosing with Vitamin C at three dose levels (0.5, 1 and 2 mg) and with each of two delivery methods (orange juice or pure ascorbic acid).

This report explores these data, through inspection of its tabular form, visualization and the construction of conference intervals hypothesis tests.

Loading and Inspection of Data

The ToothGrowth data may be loaded in R with the data function and without need to reference any libraries.

In Appendix A, we can see the types of the data and initial values using the str function. Note the two levels of the supp factor - "OJ" (orange juice) and "VC" (ascorbic acid).

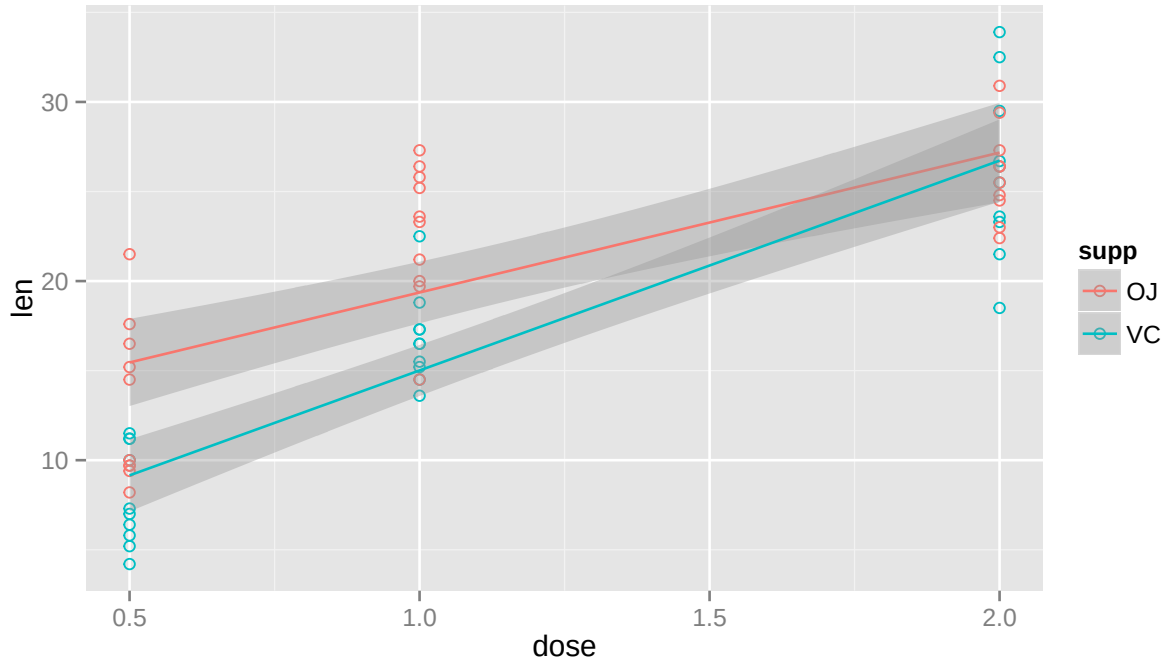
We can also see a summary of the data, including quartiles for continuous variables and counts for discrete variables, yielded by the summary function and alternatively we see the data viewed directly (using the head function to truncate output).

Visualization of Data

The data can be visualized in a scatter plot using the ggplot2 library, as can be seen in Appendix B. Color can be used to differentiate the different supplement types.

We might immediately suspect that a greater dose is associated with greater tooth length and that perhaps orange juice is a more effective delivery mechanism, at least for the smaller doses.

Indeed this becomes clearer when lines of best fit are applied as follows (code in Appendix B):



Multiplicity

If we construct intervals or accept/reject hypotheses with 95% confidence (equivalently a 5% error rate) but we perform many such tests then we can expect to see false positives. If we perform 100 tests with 95% confidence, for example, we may expect false positives in 5 of those tests.

A number of methods exist to correct for multiple testing, including the following:

- Bonferroni method (control the Family-Wise Error Rate)
- BH method (control the False Discovery Rate)

Since our number of tests will be low and for clarity of exposition we will assume multiplicity not to be a significant issue for this report but it should be noted that confidences expressed hereafter will be necessarily optimistic.

Confidence Intervals

One can construct a confidence interval in order to project a likely range for a value (here length) or to perform a comparison of factors relative to that value.

Tooth length

Z Test If we assume that n is large, i.e. large enough for the Central Limit Theorem to apply, we can calculate a confidence interval using a Z test (where SE is the standard error, $s/\sqrt{(n)}$):

$$Est \pm ZQ \times SE_{Est}$$

Thus, as Appendix C shows, [17.18900, 20.43766] is a 95% confidence interval for the length in the population. In other words, one can be 95% confident that the population mean lies within this interval (given the assumption that n is large).

T Test If we are not willing to assume n is large (and in the tooth data one may argue it is not) then we can instead use a T test. The T test may be implemented similarly to the Z test or one may use the R function t.test, as also shown in Appendix C.

Note that the interval is close but not identical to that of the Z test.

Supplement comparison

We can also construct a confidence interval comparing the two different supplement types (orange juice and ascorbic acid). As shown in Appendix C this interval is [-0.1670064, 7.5670064] and, since the interval includes 0, we cannot be sure that the true difference is not 0 and that the difference is significant at that confidence level.

Note that we assumed the populations to have equal variance (as specified by var.equal = TRUE). This is a strong assumption made here for demonstration that we will relax later.

Hypothesis Testing

One may also compare two factors using a hypothesis test. One specifies a null hypothesis, H_0, representing the status quo and an alternative hypothesis, H_a.

Dose

We may test the alternative hypotheses that there is a difference in length attributable to dose against the null hypothesis (no difference in the length between doses) in a one-sided T test, as can be found in Appendix D.

We see that the T test rejects the null hypothesis in favor of the alternative hypothesis: "true difference in means is not equal to 0".

We can also see the confidence interval and can note that it is entirely above zero. In fact analyzing the CI in this way and performing an hypothesis test are equivalent.

Note finally that here we relaxed the assumption of equal variance, yielding a narrower interval than when assuming equal variance, but in this case not affecting our results.

Supplement at low doses

Finally, we can test the hypothesis that, at lower doses (< 2.0) orange juice is a more effective supplement for than ascorbic acid.

In Appendix D we see that the null hypothesis is again rejected in favor of the alternative: that, for small doses, the mean length of the orange juice population is greater than that of the ascorbic acid population.

Summary

We have loaded and visualized the ToothGrowth dataset in R and we have created a number of confidence intervals and performed a number of hypothesis tests against it. We have discussed and explored assumptions including equality of variance, largeness of sample size and the effect of multiplicity of testing.

Appendix A - Loading and Inspection of Data

```
data(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 ...
```

```
summary(ToothGrowth)
```

```
##       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
```

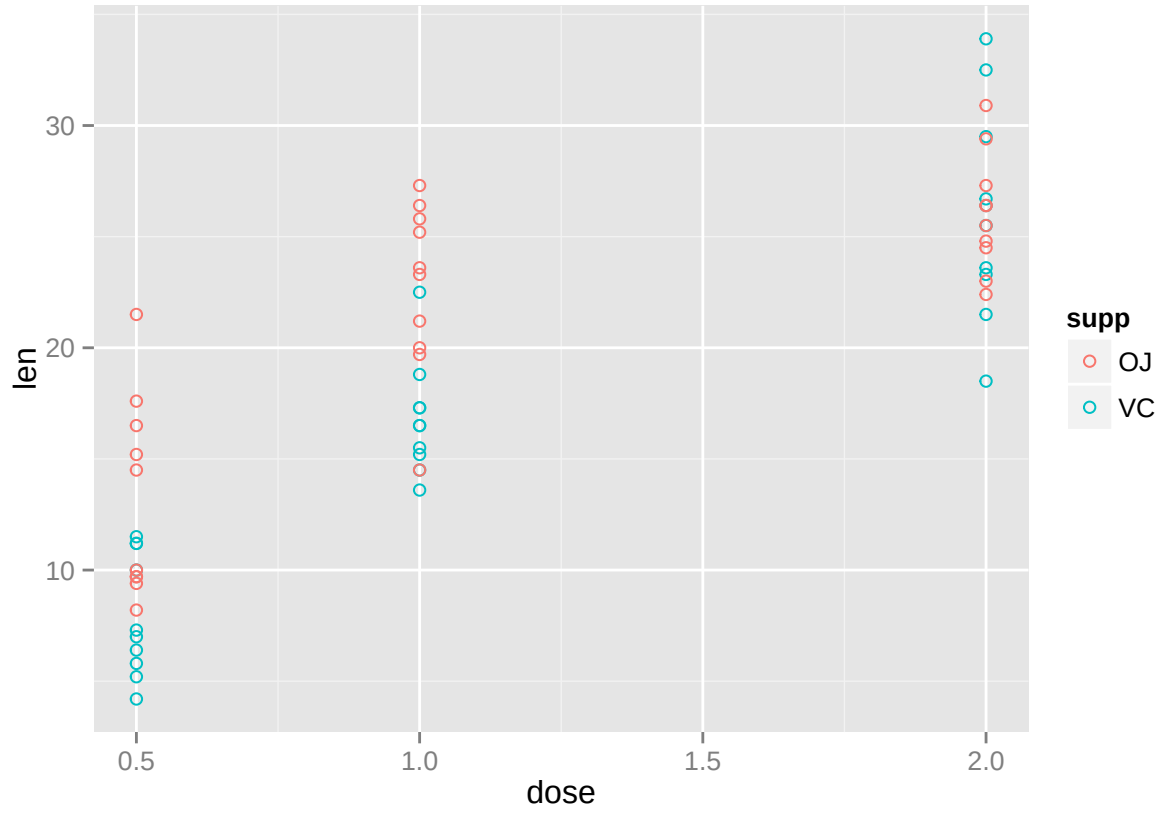
```
head(ToothGrowth)
```

```
##    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
```

Appendix B - Visualization of Data

Raw

```
library(ggplot2)
ggplot(ToothGrowth, aes(x=dose, y=len, color=supp)) + geom_point(shape=1)
```



Line of Best Fit Code

```
ggplot(ToothGrowth, aes(x=dose, y=len, color=supp)) + geom_point(shape=1) + geom_smooth(method=lm)
```

Appendix C - Confidence Intervals

Tooth Length

```
mn <- mean(ToothGrowth$len)
s <- sd(ToothGrowth$len)
z <- qnorm(0.95)
n <- nrow(ToothGrowth)
mn + c(-1, 1) * z * s / sqrt(n)
```

```
## [1] 17.18900 20.43766
```

Supplement Comparison

```
as.vector(t.test(ToothGrowth$len)$conf.int)
```

```
## [1] 16.83731 20.78936
```

```
oj <- ToothGrowth[ToothGrowth$supp == "OJ",]$len
vc <- ToothGrowth[ToothGrowth$supp == "VC",]$len
as.vector(t.test(oj, vc, var.equal = TRUE)$conf.int)
```

```
## [1] -0.1670064  7.5670064
```

Appendix D - Hypothesis Testing

Dose

```
lowDose <- ToothGrowth[ToothGrowth$dose == 0.5,]$len
highDose <- ToothGrowth[ToothGrowth$dose == 2.0,]$len
t.test(highDose, lowDose, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  highDose and lowDose
## t = 11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  12.83383 18.15617
## sample estimates:
## mean of x mean of y
##    26.100    10.605
```

Supplement at low doses

```
lowerDoses <- ToothGrowth[ToothGrowth$dose != 2.0,]
ojLower <- lowerDoses[lowerDoses$supp == "OJ",]$len
vcLower <- lowerDoses[lowerDoses$supp == "VC",]$len
t.test(ojLower, vcLower, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  ojLower and vcLower
## t = 3.0503, df = 36.553, p-value = 0.004239
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.875234 9.304766
## sample estimates:
## mean of x mean of y
##    17.965    12.375
```