R Tutorial for STAT 350 Lab 8

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

Example: (Data Set: eduproduct.txt)

Evaluation of a New Educational Product. A company designed two versions of a new educational product to improve children's reading comprehension. They would like to claim that children will acquire better reading comprehension skills utilizing either of the two versions than with the traditional approach. You are helping the company to prepare the marketing material, which will include the results of a study conducted to compare the two versions of the new product with the traditional method. The standard method is called Basal (B), and the two variations of the new method are called DRTA (D) and Strat (S).

Education researchers randomly divided 66 children into three groups of 22. Each group was taught by one of the three methods. The response variable is a measure of reading comprehension called Comp that was obtained by a test taken after the instruction was completed.

a) Make side-by-side boxplots and an effects plot of the data. Also, make a table containing the sample size, mean, and standard deviation for each treatment group. From this information, do you think that all of the means are the same? Be sure to comment on each of the plots.
b) Examine the assumptions necessary for ANOVA. Is it appropriate to continue the analysis? Be sure to state each of the assumptions including those ones that you are assuming to be true, and comment on each of them using the appropriate plots/data. Remember, you need to generate the normal probability plots and histograms for each population.
c) Report the results of the ANOVA significance test (four steps) using a significance level of 0.05. Are your results in this part consistent with part a)?
d) Use an appropriate multiple-comparison method to determine which of the educational method(s) helps reading comprehension the best. To determine which method is best, we need to compare all pairs of treatment methods. Explain why you chose this multiple-comparison method. Present a graphical representation of the results if appropriate for your method. Write a short statement for your conclusion in complete English sentences.
e) Write a short report explaining the treatment effects of this three educational methods. Be sure to answer the question posed in this question and how far the findings of this study can be generalized. This paragraph should be written in complete English sentences and should be understandable to someone who has not taken a course in Statistics.

Solution:

```
# read data
# Comp is numeric and GroupName is categorical
ed <- read.table("eduproduct.txt", header = TRUE)
```

Or, to read in the data set: Import Dataset → From CSV → Browse to find file → Delimiter: Tab, Name: "ed" → Import

R Tutorial for STAT 350 Lab 8

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

```
# When using RStudio to import your data, the program may not
#   correctly determine which variables are categorical. This is not a
#   problem in most of the functions used in the tutorials because they
#   automatically convert the class of the variables to the correct
#   one when called. However, not all functions can automatically do
#   the conversion. In that case, you need to explicitly tell R that
#   the variable is categorical by using the function as.factor(). This
#   should be done right after the data table is read in.
ed$GroupName <- as.factor(ed$GroupName)
#
# Select only rows and variables that we want
ed <- subset(ed,
             GroupName == "B" | GroupName == "D" | GroupName == "S",
             select = c("GroupName", "Comp"))
# The select argument above keeps the specified variables and removes
#   the other variables from the dataset. This is optional.
#
library(ggplot2)
# a) Side-by-side boxplot
# See Lab 7 for the code.  Remember to use x=GroupingVariable and
#   y=NumericResponseVariable
#
# a) Effects plot
# We use two calls to stat_summary(): One plots points at the sample
#   means of each category, the other connects the points with lines
windows()
ggplot(data = ed, aes(x = GroupName, y = Comp)) +
  stat_summary(fun.y = mean, geom = "point") +
  stat_summary(fun.y = mean, geom = "line", aes(group = 1)) +
  ggtitle("Effects Plot of Comp by GroupName")
# a) Calculating the sample size, sample mean and standard deviation
tapply(ed$Comp, ed$GroupName, length)
tapply(ed$Comp, ed$GroupName, mean)
tapply(ed$Comp, ed$GroupName, sd)
#
# b) Histogram
# A little more complicated because we have three groups now. The code
#   consists of two steps.
#
# (1) Make theoretical density curve
# You need to specify the name of the numeric response in the square
#   brackets of as.numeric(x[]),
#   and the name of the grouping variable in xbar[x[]] and s[x[]].
#
xbar <- tapply(ed$Comp, ed$GroupName, mean)
s    <- tapply(ed$Comp, ed$GroupName, sd)

ed$normal.density <- apply(ed, 1, function(x){
  dnorm(as.numeric(x["Comp"]),
        xbar[x["GroupName"]], s[x["GroupName"]])})
```

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

```
#
# (2) Make the histogram
#
windows()
ggplot(ed, aes(x = Comp)) +
  geom_histogram(aes(y = ..density..), bins = sqrt(nrow(ed)) + 2,
                 fill = "grey", col = "black") +
  facet_grid(GroupName ~ .) +
  geom_density(col = "red", lwd = 1) +
  geom_line(aes(y = normal.density), col = "blue", lwd = 1) +
  ggtitle("Histograms of Comp by GroupName")
#
# QQ plot
#
# (1) Calculate slope and intercept
# Need to put name of the grouping variable in xbar[x[]] and s[x[]].
#
ed$intercept <- apply(ed, 1, function(x){xbar[x["GroupName"]]})
ed$slope <- apply(ed, 1, function(x){s[x["GroupName"]]})
#
# (2) Make the QQ plot
#
windows()
ggplot(ed, aes(sample = Comp)) +
  stat_qq() +
  facet_grid(GroupName ~ .) +
  geom_abline(data = ed, aes(intercept = intercept, slope = slope)) +
  ggtitle("QQ Plots of Comp by GroupName")
#
# ANOVA
# c) hypothesis test
# The command is aov(numeric ~ categorical, data = datasetName)
# The function is called aov(): a as in analysis, o as in of, and
#   v as in variance.
# The categorical variable must be of the class "factor" or "integer."
# To print out the results, you need to use the function summary().
#   Note: this does not print out the "total" line in the ANOVA table.
#   you may calculate it by hand or via R if required.
fit <- aov(Comp ~ GroupName, data = ed)
summary(fit)
# d) Tukey method
test.Tukey <- TukeyHSD(fit, conf.level = 0.95)
test.Tukey
```

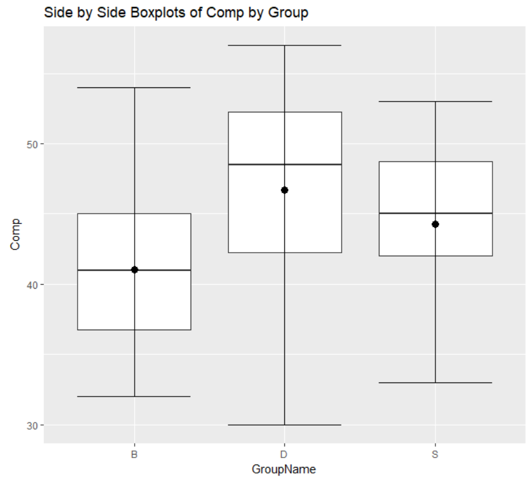
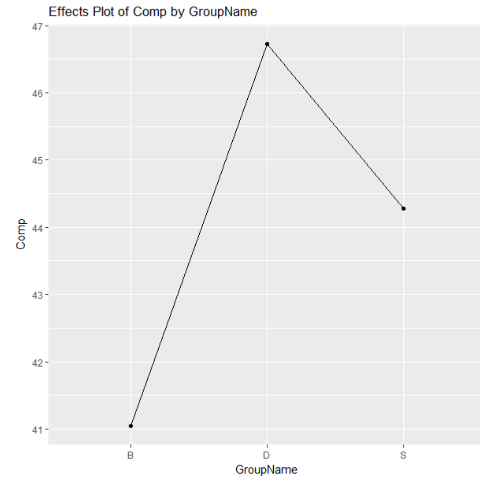
Remember to use the `as.factor()` function if R doesn't recognize that your categorical variable is a categorical variable.

a) Make side-by-side boxplots and an effects plot of the data. Also, make a table containing the sample size, mean, and standard deviation for each treatment group. From this information, do you think that all of the means are the same? Be sure to comment on each of the plots.

R Tutorial for STAT 350 Lab 8

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

Solution:

<table>
<tr><th>side-by-side boxplots</th><th>effects plot</th></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>From this plot, I would state that all of the means could be the same, since the boxes overlap a fair amount.</td><td>From the effects plot, we can get a sense of the ranking of the means, if they turn out to be statistically different.</td></tr>
</table>

Calculating the sample size, mean and standard deviation per category:

sample size:

```
> tapply(ed$Comp, ed$GroupName, length)
 B  D  S 
22 22 22 
```

mean:

```
> tapply(ed$Comp, ed$GroupName, mean)
       B        D        S 
41.04545 46.72727 44.27273 
```

standard deviation:

```
> tapply(ed$Comp, ed$GroupName, sd)
       B        D        S 
5.635578 7.388420 5.766750 
```

Group	n	sample mean	sample standard deviation
B	22	41.04545	5.635578
D	22	46.72727	7.388420
S	22	44.27273	5.766750

It appears group B may have a smaller mean than the other two, but inference needs to be performed (both c) and d)) to determine if B is lower than the other treatment groups or they are all the same.

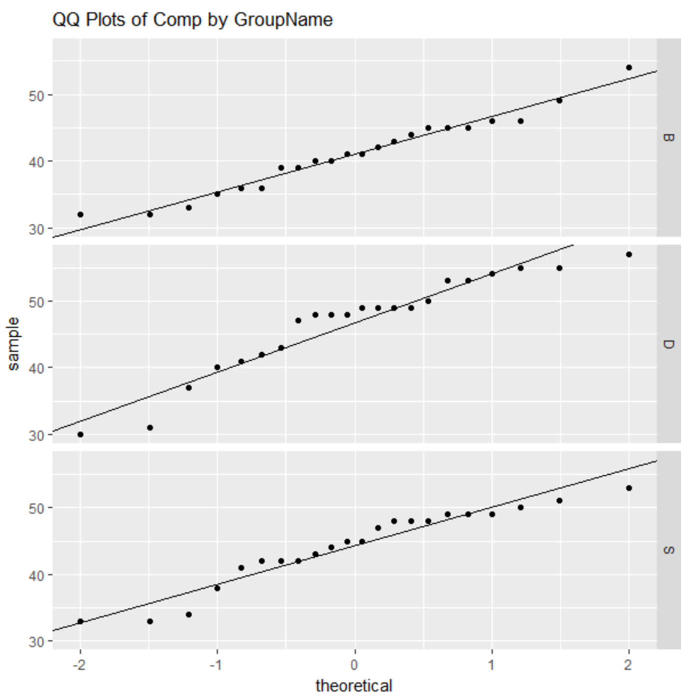
Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

b) Examine the assumptions necessary for ANOVA. Is it appropriate to continue the analysis? Be sure to state each of the assumptions including those ones that you are assuming to be true, and comment on each of them using the appropriate plots/data. Remember, you need to generate the normal probability plots and histograms for each population.

Solution:

SRS / independent errors:
Assumed to be true.

Normality:



At the given sample size, these distributions are close enough to being normal.

Constant standard deviation

GroupName	n	sample mean	sample standard deviation
B	22	41.04545	5.635578
D	22	46.72727	7.388420
S	22	44.27273	5.766750

$$\frac{s_{max}}{s_{min}} = \frac{7.388420}{5.635578} = 1.31 < 2$$

Therefore, the constant standard deviation assumption is valid.

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

Because all of the assumptions are valid, it is appropriate to continue the analysis.

c) Report the results of the ANOVA significance test (four steps) using a significance level of 0.05. Are your results in this part consistent with part a)?

Solution:

```
             Df Sum Sq Mean Sq F value Pr(>F)  
GroupName     2  357.3  178.65   4.481 0.0152 *
Residuals    63 2511.7   39.87                 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 1: Define parameters

μ_B is the population mean Comp score for the Basal method (control).
μ_D is the population mean Comp score for the DRTA method.
μ_S is the population mean Comp score for the Strat method.

Step 2: State the hypotheses

H_0: $\mu_B = \mu_D = \mu_S$
H_a: at least two μ_i's are different.

Step 3: Find the *test statistic, p-value, report DF*

$F_{ts} = 4.481$
DF1 = 2, DF2 = 63
p-value = 0.0152

Step 4: Conclusion:

$\alpha = 0.05$
Since 0.0152 < 0.05, we should reject H_0

The data provides evidence (p-value = 0.0152) to the claim that the population mean values of at least one of the education methods is different from the rest.

In part a), I stated that the methods might have been different, but I wasn't sure because the three groups looked potentially the same from the boxplot. Here, the test based on ANOVA indicated a difference. Therefore, the conclusions are partially consistent. Keep in mind that the results of the test are more objective than the subjective analysis of the plots.

d) Use an appropriate multiple-comparison method to determine which of the educational method(s) helps reading comprehension the best. To determine which method is best, we need to compare all pairs of treatment methods. Explain why you chose this multiple-comparison method. Present a graphical representation of the results if appropriate for your method. Write a short statement for your conclusion in complete English sentences.

R Tutorial for STAT 350 Lab 8

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

Solution:

The Tukey method was chosen because we want to compare all of the means in a pairwise fashion.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Comp ~ GroupName, data = ed)

$GroupName
         diff        lwr       upr     p adj
D-B  5.681818  1.112137 10.251499 0.0111135
S-B  3.227273 -1.342408  7.796953 0.2149995
S-D -2.454545 -7.024226  2.115135 0.4064363
```

There are two ways to determine which pairs are significant.
1) Check whether 0 is in the interval (`lwr`, `upr`)
2) Check whether the `p adj` is smaller than the significance level.

In this case, we have statistical evidence that D and B are different but we do not have evidence that B is different from S nor that D is different from S.

The following is the graphical representation (to determine the values of the means, please look at the table that you created in part a). You must construct this by hand (unless you want to program R to do this by yourself).

B	S	D
41.04545	44.27273	46.72727

Note that the above is a table was created in Word where I used the Border Painter in Table Tools → Design to only include the appropriate Borders.

This says that that, in terms of statistical significance, methods D and S are the same and methods S and B are the same (because they are joined by an underline). Since we are interested in the method that leads to the largest value, we should eliminate method B because it is significantly different from method D: The second underline only joins S and D. Meanwhile, note that method S is not significantly different from the worst method, B. Therefore, I would choose educational method D out of S and D.

e) Write a short report explaining the treatment effects of this three educational methods. Be sure to answer the question posed in this question and how far the findings of this study can be generalized. This paragraph should be written in complete English sentences and should be understandable to someone who has not taken a course in Statistics.

R Tutorial for STAT 350 Lab 8

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

Solution:

The original question asked us to determine the best educational method and whether the new methods D and S are better than the traditional method, B. We determined that the ANOVA assumptions were valid, so we performed an ANOVA analysis to infer about the population reading comprehension scores, Comp, associated with these methods. These results show that method S is not statistically different from method B (traditional method). However, we have evidence that method D is better than the original method. The study can be generalized to other children who are at similar levels of reading comprehension as the children in the study.

Remember when discussing the conclusion in other situations, first determine whether you are interested in the lowest number(s) or the highest number(s).