Stat-300 – Final Exam

Name

SOLUTIONS

The model solutions are comprehensive – I didn't expect that students would respond with all of the points present in the solutions; nor did I expect the students to write as much as in my model solutions.

Student Number:

Put your name and student number on the upper right of each of the following pages in case the pages get separated.

Answer the following questions in the space provided. Be sure that your answers are legible. The majority of the marks is allocated to organization and lucidity - pitch your answers to someone with a single course in Statistics.

1. Time is Money - I - 5 marks.

Please explain the sentences below. This quote is taken from the *Time is Money* article.

> The extent of cheating also varied across conditions, $F(2, 95) = 5.09$, $p = .008$. Simple contrasts revealed that participants cheated more in the money condition ($M = 4.41$, $SD = 4.25$) than in both the control condition ($M = 2.76$, $SD = 3.96$; $p = .07$) and the time condition ($M = 1.55$, $SD = 2.41$; $p = .002$). The difference between the time and control conditions did not reach statistical significance ($p = .18$).

Solution:

An experiment was conducted to measure the amount of cheating among students randomly assigned to three different treatment groups (denoted *time*, *money* and *control*). A single factor completely-randomized ANOVA found evidence that the mean amount of cheating varied among the treatment groups ($p = .008$). This implies that the differences in the observed mean response between the three groups was larger than would have been expected if the three groups had the same average amount of cheating.

The ANOVA does not provide information on where the differences in mean could be, i.e. the result of the ANOVA just indicates that there is evidence that not all three means are equal. The researcher then did simple pairwise comparisons, i.e. examined if there was evidence of a difference in the means of each pair of treatment groups.

They claimed that they found evidence that the mean amount of cheating was larger in the *money* treatment group compared to the *control* treatment group (estimated difference in the means is 1.65 (SE xxxx)[1], $p = .07$), but the largish p-value actually shows that the evidence is very weak and that the observed difference in the means is not that unusual relative to what would be expected had the means actually been the same.

The authors found evidence that the mean amount of cheating was different between the *money* and *time* treatment groups (est diff 2.86 (SE xxxx), $p = .002$), but found no evidence of a difference the mean amount of cheating between the *time* and *control* treatment groups (est diff 1.21 (SE xxxx), $p = 0.18$).

Common Errors made by students

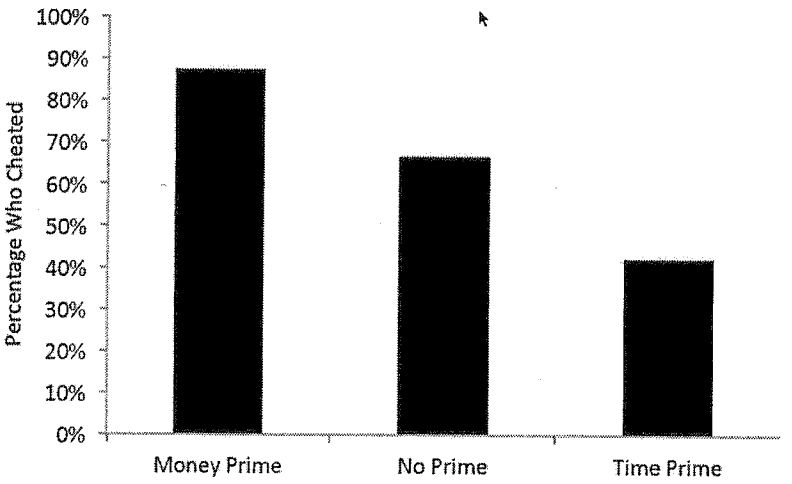
- Students forgot to express the conclusions in terms of the MEAN amount of cheating, i.e. simply said that the amount of cheating varied among the groups.
- Students did not express in terms of *evidence*, but made statement such as "The authors proved that a difference existed."

[1] You can't compute a SE here because not enough information was given in the problem

2. Time is Money - II - 5 marks

Consider the following graph taken from the *Time is Money* article that we studied in one of the assignments.

Figure 1. Percent of participants who cheated in Experiment 1 by condition



The authors concluded:

> The percentage of participants who cheated varied across conditions, $\chi^2(2, N = 98) = 14.61$, $p = .001$ (see Figure 1); participants were more likely to cheat in the money condition (87.5%, 28/32) than in either the control condition (66.7%, 22/33; $\chi^2(1, N = 65) = 3.97$, $p < .05$) or the time condition (42.4%, 14/33; $\chi^2(1, N = 65) = 14.44$, $p < .001$). Participants were also less likely to cheat in the time condition than in the control condition ($\chi^2(1, N = 66) = 3.91$, $p < .05$).

WITHOUT DOING ANY COMPUTATIONS, modify the graph above to reflect these conclusions. Explain how your modification reflect these conclusions? Also, how would you improve reporting the results?

Solutions:

(a) You need to add confidence interval "error" bars to the chart, but the confidence intervals have NO OVERLAP between them across the three bars. The screen of the bars should also be lighter so that the upper and lower extent of the confidence intervals can be clearly seen. Because the proportion who cheat + the proportion who do not cheat must add to 100%, a segmented bar chart (with the confidence interval overlaid) could also be done.

(b) There are several things that can be improved in the reporting.

First, it is unnecessary to report the actual percentages to 2 decimals places as the small sample sizes implies that the standard errors are sufficiently large that you would barely know the first digit never mind the 4th significant digit. The percentages should be rounded to integers.

Second it is bad form to report p-values merely as $p < 0.05$. The exact p-value should be reported so that the reader can judge if the p-value is 0.049 or 0.001. Of course this must be reasonable; you should not report a p-value of .00000001 as this makes many (unwarranted) assumptions about the distribution of the data values (e.g. all value are independent) and the sampling distribution of the test statistic (e.g. is the chi-square approximation valid?)

Third, statistical significance does not imply that there is a difference in the population parameters (Type I or false positive results are possible). The conclusions should be couched as "There was evidence that the percentage of participants who cheated varied across conditions" etc.

Fourth, never report a naked estimate. The estimated proportion of people who cheated in each condition should have a measure of precision reported, either a standard error or a 95% confidence interval.

Lastly, it appears that they made no corrections for multiple-comparisons and did simple contrasts after the omnibus test reported in the first line. In order to control the overall (familywise) false positive rate, the authors need to use an adjustment (such as a Tukey-like adjustment) when making the pairwise comparisons.

Common Errors made by students

- Students tried to fit a regression line to the char. The treatment categories are nominal scaled and so this is not appropriate.
- Students suggest than an ANOVA should be used ANOVAs are used when comparing means among groups – this article is comparing proportions and so the use of a chi-square test (or a form of logistic regression) would be appropriate – assuming, of course, that the analysis was appropriate for the experimental design.

3. Did you enjoy the Tukey at Thanksgiving - 5 marks

Dear Stat Guru:

I'm doing a study comparing egg sizes (mm) of cuckoo eggs from different host-species of birds. I did a Tukey-procedure following an ANOVA and got the following results:

Level			Least Sq Mean
Hedge	A		23.121429
Tree	A		23.090000
Wagtail	A	B	22.903333
Robin	A	B	22.575000
Meadow		B	22.298889
Wren		C	21.130000

Levels not connected by same letter are significantly different.

There are a couple of things I don't understand about this output:
(a) Why do I need a Tukey-procedure?
(b) What is the above telling me?
Thanks, a confused client.

Solution: (a) The overall ANOVA test only indicates that there is evidence that at least one group's mean differs from the other group means, but provides no information on where the differences in means may lie. Following the ANOVA, a further analysis is done where the means from each of the possible pairs of groups (15 possible pairs) is examined to see if there is evidence that they are unequal.

Unfortunately, a naive application of, for example, a simple t-test leads to problems. Each of the 15 possible pairs had a 5% chance of a false positive. The overall chance (over all 15 comparisons) is then much larger than 5%. Consequently, the chance of a false positive somewhere among the 15 possible comparisons is much larger than 5%. The Tukey procedure (and many other multiple comparison procedures) ensure that the overall false positive error rate over all of the 15 possible comparisons is controlled to be 5% or less by restricting the false positive rate for the individual comparisons to a level less than 5%.

(b) The jointed-letters display (a.k.a. compact-letter display or joined-line display) is way to present the results of the 15 pairwise comparisons in a compact fashion. Consider any pair of means. If these two groups are "joined" by the same letter, then there is no evidence that the means of the two groups are different. The actual letter code has no particular meaning. In this case, because the Hedge mean and Tree mean are joined by the letter A, there is no evidence that their respective population means differ. However, both of these groups are NOT joined to the Wren group. We conclude that there there is evidence that both means differ from the mean of the Wren group.

Notice that these comparison are NOT transitive. For example, There is no evidence that the mean for Hedge is different from the mean for Robin (both are joined by the letter A); there is no evidence that the mean of Robin is different from the mean for Meadow (both joined by the letter B). Yet, the mean for Hedge is not joined to the mean for Wren by the same letter. The can be explained by the paint-chip analogy used in class – three paint shades are compared. A side-by-side comparison of the first and second chip is unable to detect a difference; between the second and third chip again is unable to detect a difference; but the first and third chip are sufficiently different that they can be distinguished.

Common Errors made by students

- Students did not indicate that the ANOVA and Tukey-comparisons are about MEANS.
- Students did not explain that if treatments groups are joined by the same letter, there is no evidence of a difference in the MEAN egg size.

4. Water quality sampling - 5 marks

A survey was done to compare water turbidity (the opposite of clarity; measured in NTU) at five different sites along French Creek, a stream on Vancouver Island. The five sites are labelled as BB, Coombs, Grafton, NewHwy, and WinchRd.

Samples were taken at 15 different sampling times (denoted as YYYY.MMDD, where YYYY year, MM=Month, and DD=Date). Because of budget constraints, only three of the five sites were measured at each sampling time.

What is the technical term for this design? What advantages does this design have compared to a design where a single stream is measured on different days (for a total of 45 sampling days with each stream measured on 9 different days and no pairs of streams measured on the same day). What potential problems do you see with the fact that only 3 sites are measured on each day?

Solution:

This is an (incomplete) blocked-design where the sampling date is the blocking variable.

The blocked design is used to account for common causes that may cause the turbidity to co-vary among the streams together. For example, all streams would experience the spring melt at roughly the same time and so the turbidity in all streams may be higher because they are all experiencing the freshet. By blocking by date, it is possible to control for these common causes (differences within blocks will be free of block effects).

By blocking by date, some cost saving could occur because, presumably, there is less set up time each day. The total costs for the analysis of the NTU is still the same (45 measurements will still be taken).

The design is incomplete because not all streams were measured on each sampling date. This causes NO undue complications in the analysis with modern software, unless the pattern of sampling was such that the design was NOT connected. For example, suppose that streams BB, Coombs, and Grafton are always measured on the same dates in the spring, and the streams NewHwy and WinchRd are always measured on the same dates in the fall. Then differences in the mean turbidity between these two groups of streams is confounded with block effects and/or seasonal effects.

The analysis of an incomplete block design has two parts – the recover of the intra-block (within block) information and the recovery of the inter-block (among block) information. The intra-block analysis treats blocks as fixed effects; the recovery of both types of information requires that block effect be random effects from a larger population of blocks.

Common Errors made by students

- Students said that blocking REMOVES the effect of date. It is still there, but blocking controls FOR date effects when the intra-block comparisons are made.

5. Sample size - 5 marks

Explain how the standard deviation and desired margin of error influence the required sample size for a survey. Explain the influence of population size (number of units in the population) on the required sample size for surveys.

Solution:
The required sample size for a survey depends on several factors - two of which are the standard deviation in the response values and desired margin of error.

- The standard deviation in the response. A larger standard deviation implies that a larger sample size will be needed, *ceteris paribus*, to obtain estimates with a desired precision compared to a smaller standard deviation. At an extreme, if the standard deviation was 0, then all units have the identical response, and a sample size of 1 gives perfect information!
- The margin of error (typically defined as twice the standard error) is the desired uncertainty in the estimate derived from the survey. Every different survey, because it is based on a sample from the population, will give a different estimate. The reproducibility of the estimate across different samples is measured by the standard error (and the margin of error). If an estimate is to have a very small standard error, then estimates from different samples must be highly reproducible and will require a much larger sample size to ensure that minor changes in data values between different sample, do not influence the final estimate very much.

Surprisingly, the population size has little influence on the required sample size for reasonable surveys, i.e. where the number of units sampled is much smaller than the population size. The analogy of the pot-of-stew explained in class is directly applicable.

Suppose you are cooking a 1 L pot of stew. You wish to see if the peas in the stew are cooked. You would stir the stew (randomize) and sample a small amount (say 25 mL). Now suppose you are cooking a 10 L pot of stew. You again wish to see if the peas in the stew are cooked. You again stir (randomize), but would again only sample a small amount (25 mL). Just because the second pot is 10× larger than the first pot, you would not increase your sample size to 250 mL!

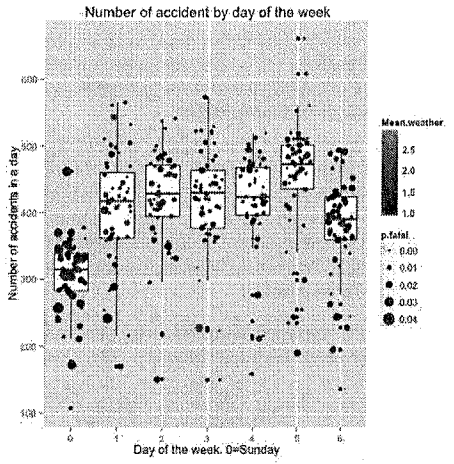
Common Errors made by students

- Students got confused between the SD and the SE (or margin of error).

6. Road accidents in 2010 - 5 marks

Consider the following plot showing the relationship between the day of the week (0=Sunday), number of accidents in day, the proportion of accidents that had a fatality (p.fatal), and the mean weather severity (mean.weather) where higher numbers are associated with more severe weather (e.g. snow is treated as more severe than rain which is treated as more severe than fair):

Note that students were given a full size blowup of the plot in colour during the exam.



Create a short writeup summarizing the results; explain how the plots illustrates your results.

The plot is showing the relationship between the number of accidents in day involving personal injury, the proportion of fatalities, the day of the week, and a measure of weather severity.

Consider first the relationship between the number of accidents and the day-of-the week. Not surprisingly the distribution of the number of accidents on a Sunday (day-of-the-week 0) is shifted downwards compared to the Monday to Friday and even Saturday. The distribution of the number of accidents on Friday appears to have a small shift upwards while that on Saturday appears to have a small shift downward compared to Monday to Thursday. This is summarized by the boxes and whiskers (the box plots). The central box of the box-plot shows the central 50% of the distribution for the number of accidents/day. It is quite evident from the plot that these boxes are shifted upwards and downwards.

The proportion of accidents with a fatality in a day is indicated by the size of the dot with larger dots corresponding to a higher proportion of fatalities. Somewhat surprising to me is that while Sunday appears to have a smaller number of accidents per day, the proportion of fatalities tends to be higher than other days. This is seen by the preponderance of large dots on Sundays compared to the other days of the week. Saturday also seems to have higher proportions of fatalities, but not nearly as much as Sundays.

The relationship of number of accidents and proportion of fatalities with weather severity is harder to discern. This is encoded using the colour of the dot, with lighter colours representing more severe weather. We can see a few dots that are light blue (severe weather), but have lower numbers of accidents and intermediate fatality rates (the dots are not that large), and a few light blue dots (indicating very severe weather) with surprising low proportion of fatalities (small size of dotes). However, it is not clear at all from this graph what the relationship is between weather severity and the other variables.

Common Errors made by students

- Students treated the box-plots as confidence intervals.

7. All I want for Christmas are my 2 AAs - 10 marks

Does it make a difference of what brand of battery is used to power children (or adult) toys? Think of a way to run such an experiment as a CRD. Write a *Materials and Methods* for your thought experiment. Use the attached computer output and write a *Results* section.

Use this and the next page for your answer. (Hint: put the M & M on this page and the Results on the next page.)

Solution: *Material and Methods:*
A selection of batteries from four brands was obtained from local supermarkets and drug stores. We tried to ensure that all batteries were fresh by checking the expiration dates, and selected only a few batteries from each store. We tried to randomize which batteries were selected from each store by not simply choosing the top battery in each display rack.

A radio-controlled toy was used to measure the lifetime of each battery when used in the toy. The order in which the brands were tested was randomized. The same type of activity was used to measure the lifetime of each trial (hours).

The mean lifetime of the grands was compared using a single-factor CRD ANOVA followed by a Tukey multiple comparison procedure to compare the mean lifetime among the pairs of brands.

Results
A total of 11 trials was conducted with between 2 and 4 trials per brand (Table 1). Side-by-side box plots were used to check for outliers – non were found. The standard deviations of the lifetime (Table 1) showed that these were comparable and so the assumptions of a single-factor CRD ANOVA appear to be satisfied

There was strong evidence of a difference in the mean lifetime ($F_{3,7} = 35.1, p = .0001$). The Tukey multiple comparison procedure indicated that there was evidence that the mean lifetime for Brand 2 is larger than all the other brands, but there was no evidence of a difference in the mean lifetime among brands 1, 3 and 4 (Figure 1).

Table 1. Summary statistics from the experiment

Brand	n	Mean (hrs)	SD (hrs)
Brand 1	2	5.3	.35
Brand 2	4	7.0	.41
Brand 3	3	4.0	.50
Brand 4	2	4.3	.36

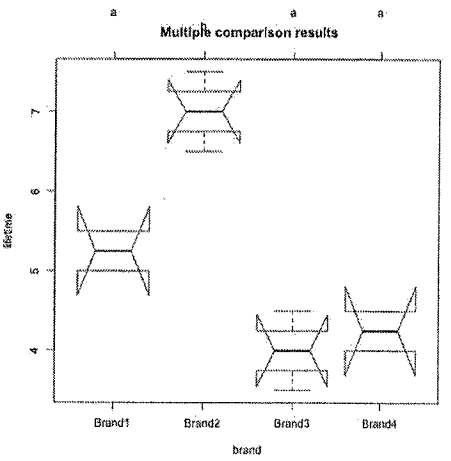


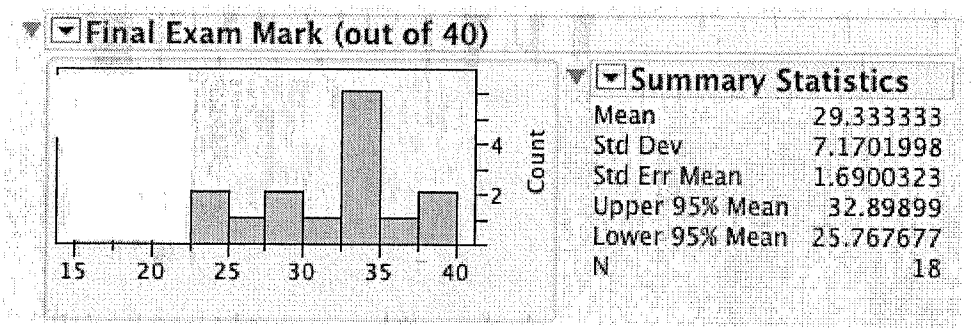
Figure 1. Estimated mean battery lifetimes by brand with 95% confidence interval (approximate notches on box plots) and compact-letter display summarizing the results of the Tukey multiple comparison procedure (letters along the top). If the notches overlap between brands, there is no evidence of a difference in mean lifetime. This is also indicated in the compact-letter-display where the brands are "joined" by the same letter.

You might be interested to know that this question is based on real life experiment in the Schwarz household.

Common Errors made by students

- Randomize, randomize, randomize. Students typically forgot to mention that the order of testing the brands should also be randomized.
- Just saying that ANOVA was used is not sufficient, You need to specify the type of ANOVA, e.g. a single-factor CRD ANOVA.
- Results should be about MEAN lifetimes.
- Table and Figure legends should be complete.

Summary Statistics about the final exam.



Some grades have been hidden.

battery.r

cschwarz — *Nov 20, 2013, 9:57 PM*

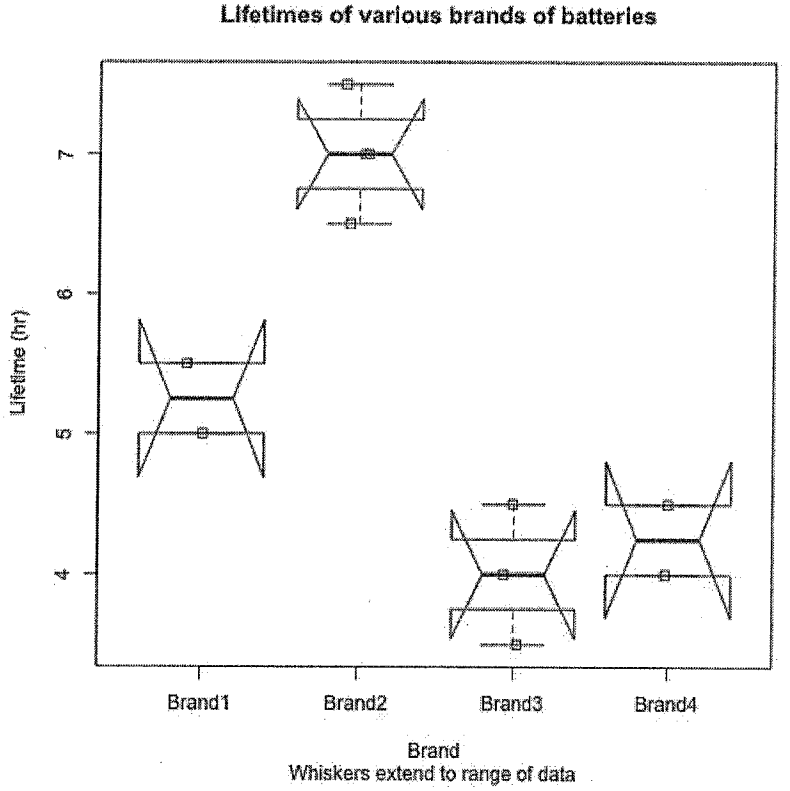
```
# Lifetime of various battery brands
options(useFancyQuotes=FALSE) # renders summary output corrects


# Read in the data
fun <- read.csv('battery.csv', header=TRUE)
fun
```

```
    brand lifetime
1  Brand1      5.5
2  Brand1      5.0
3  Brand2      7.0
4  Brand2      7.5
5  Brand2      6.5
6  Brand2      7.0
7  Brand3      4.0
8  Brand3      3.5
9  Brand3      4.5
10 Brand4      4.5
11 Brand4      4.0
```

```
# Get side-by-side dot plots
boxplot( lifetime ~ brand, data=fun, range=0,
   notch=TRUE, # helps to compare pop means
   main="Lifetimes of various brands of batteries",
   sub='Whiskers extend to range of data',
   xlab='Brand', ylab='Lifetime (hr)')
```

```
Warning: some notches went outside hinges ('box'): maybe set
notch=FALSE
```

```
stripchart(lifetime ~ brand, data=fun, add=TRUE,
     vertical=TRUE, method="jitter", jitter=.1)
```




```
# There is no easy easy way to make a nice report showing summary
# statistics by group in R unless you install some of the packages
# such as the doBy package. We will code this by hand.
# Compute some summary statistics for each group
library(doBy)
```

```
Loading required package: multcomp
Loading required package: mvtnorm
Loading required package: survival
Loading required package: splines
Loading required package: MASS
```

```
report<- summaryBy(lifetime ~ brand, data=fun,
FUN=c(length,mean,sd))
report$lifetime.se <-
report$lifetime.sd/sqrt(report$lifetime.length)
report
```

```
   brand lifetime.length lifetime.mean lifetime.sd lifetime.se
1 Brand1               2          5.25      0.3536      0.2500
2 Brand2               4          7.00      0.4082      0.2041
3 Brand3               3          4.00      0.5000      0.2887
4 Brand4               2          4.25      0.3536      0.2500
```

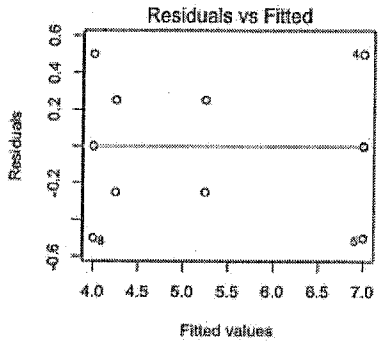
```
# get the individual confidence intervals
ci <- tapply(fun$lifetime, fun$brand, FUN=t.test)
# use the sapply() function to extract elements from each member of
the list
sapply(ci,"[","conf.int")
```

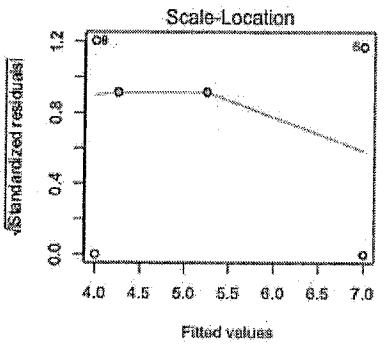
```
$Brand1.conf.int
[1] 2.073 8.427
attr(,"conf.level")
[1] 0.95

$Brand2.conf.int
[1] 6.35 7.65
attr(,"conf.level")
[1] 0.95

$Brand3.conf.int
[1] 2.758 5.242
attr(,"conf.level")
[1] 0.95

$Brand4.conf.int
[1] 1.073 7.427
attr(,"conf.level")
[1] 0.95
```

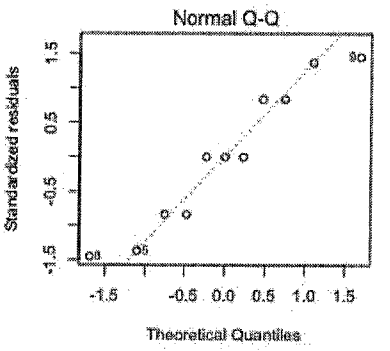
```
# fit the linear model and get the ANOVA table and test for effects
result <- aov(lifetime ~ brand, data=fun)
anova(result)
```

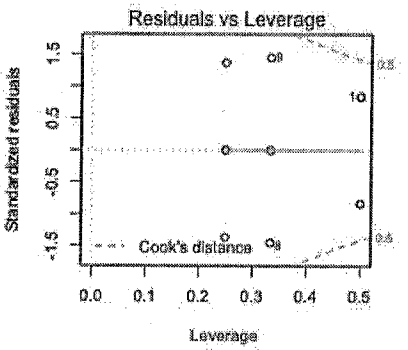
```
Analysis of Variance Table

Response: lifetime
          Df Sum Sq Mean Sq F value  Pr(>F)
brand      3  18.80    6.27    35.1 0.00014 ***
Residuals  7   1.25    0.18
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Check the assumptions of the ANOVA model
layout(matrix(1:4, 2,2))
plot(result)
```










```
layout(1)


# Estimate the marginal means.
# This is a real pain in R as you need to first get a list
# of the factor combinations that you want (the unique() function)
library(lsmeans)
```

```
Loading required package: plyr
```

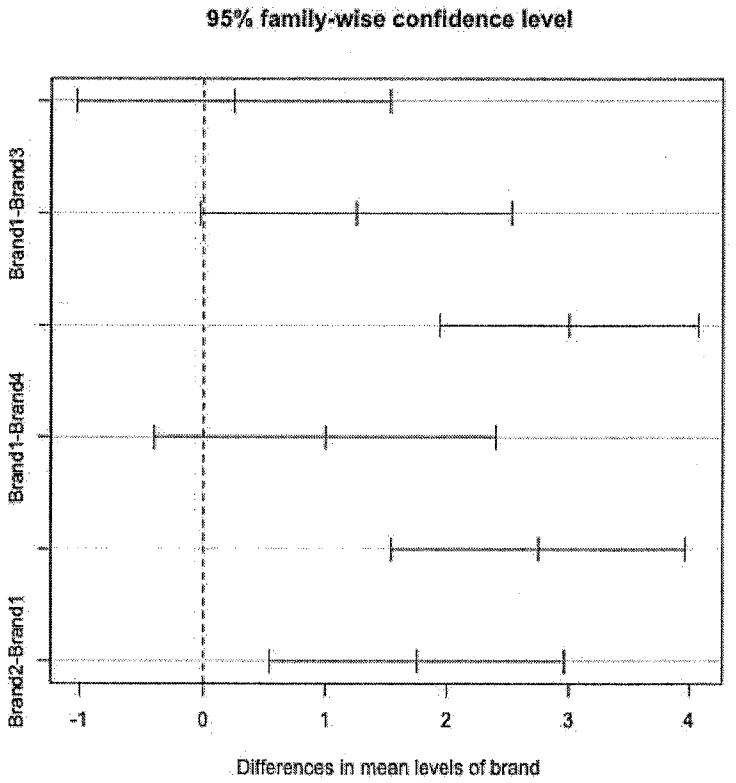
```
my.lsmeans <- lsmeans(result, pairwise ~ brand)
my.lsmeans
```

```
$`brand lsmeans`
  brand lsmean     SE df lower.CL upper.CL
 Brand1   5.25 0.2988  7    4.543    5.957
 Brand2   7.00 0.2113  7    6.500    7.500
 Brand3   4.00 0.2440  7    3.423    4.577
 Brand4   4.25 0.2988  7    3.543    4.957

$`brand pairwise differences`
                estimate     SE df t.ratio p.value
Brand1 - Brand2    -1.75 0.3660  7 -4.7819 0.00838
Brand1 - Brand3     1.25 0.3858  7  3.2404 0.05473
Brand1 - Brand4     1.00 0.4226  7  2.3664 0.17170
Brand2 - Brand3     3.00 0.3227  7  9.2952 0.00015
Brand2 - Brand4     2.75 0.3660  7  7.5144 0.00059
Brand3 - Brand4    -0.25 0.3858  7 -0.6481 0.91296
    p values are adjusted using the tukey method for 4 means
```

```
# Now for a multiple comparison procedures
mcp <- TukeyHSD(result, ordered=TRUE) # ordered sorts means
mcp
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = lifetime ~ brand, data = fun)

$brand
                diff       lwr   upr  p adj
Brand4-Brand3 0.25 -1.02692 1.527 0.9130
Brand1-Brand3 1.25 -0.02692 2.527 0.0547
Brand2-Brand3 3.00  1.93165 4.068 0.0002
Brand1-Brand4 1.00 -0.39880 2.399 0.1717
Brand2-Brand4 2.75  1.53860 3.961 0.0006
Brand2-Brand1 1.75  0.53860 2.961 0.0084
```

```
plot(mcp)
abline(v=0, lty=2)
```



```
# The multiple comparison package has lots of good routines.
# specify all pair-wise comparisons among levels of variable
"tension"
# This is a bit of R magic -- see the help pages and examples of
the packages
# for details
library(multcomp)

result.tukey <- glht(result, linfct = mcp(brand = "Tukey"))
result.tukey.cld <- cld(result.tukey)  # joined line plot

# create the display
result.tukey.cld
```

```
Brand1 Brand2 Brand3 Brand4
   "a"    "b"    "a"    "a"
```

```
plot(result.tukey.cld, main="Multiple comparison results",
     xlab="brand",
     ylab="lifetime",
     mai=c(1,1,1.25,1) , notch=TRUE)
```

```
Warning: some notches went outside hinges ('box'): maybe set
notch=FALSE
```