CODE

Statistical Computations

Joseph S. Tabadero, Jr.

November 1, 2017

Import Datasets

CODE

I have stored the data sets in the `data` folder.

CODE

Show 10 entries

Search:

	Day	Cell	Voltage	fruit
1	1	1	0.2	banana
2	2	1	0.33	banana
3	3	1	0.42	banana
4	4	1	0.65	banana
5	5	1	0.89	banana
6	6	1	0.95	banana
7	7	1	1.02	banana
8	1	2	0.22	banana
9	2	2	0.37	banana
10	3	2	0.43	banana

Showing 1 to 10 of 63 entries

Previous 1 2 3 4 5 6

Exploratory Data Analysis

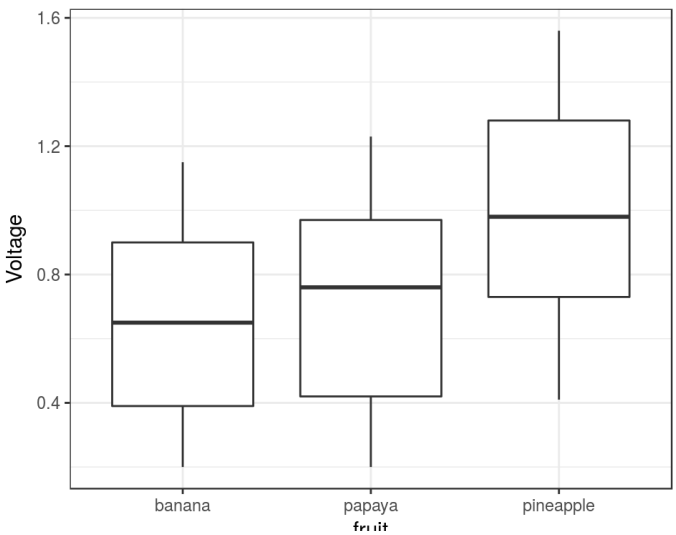
Get mean by fruit and day

CODE

```
## # A tibble: 21 x 3
## # Groups:   Day [?]
##       Day     fruit Voltage
##    <int>     <chr>   <dbl>
##  1     1    banana  0.2300
##  2     1    papaya  0.2700
##  3     1 pineapple  0.4467
##  4     2    banana  0.3633
##  5     2    papaya  0.3967
##  6     2 pineapple  0.6800
##  7     3    banana  0.4467
##  8     3    papaya  0.4933
##  9     3 pineapple  0.7900
## 10     4    banana  0.6600
## # ... with 11 more rows
```

Get the boxplots of voltage outputs by type of fruit peelings

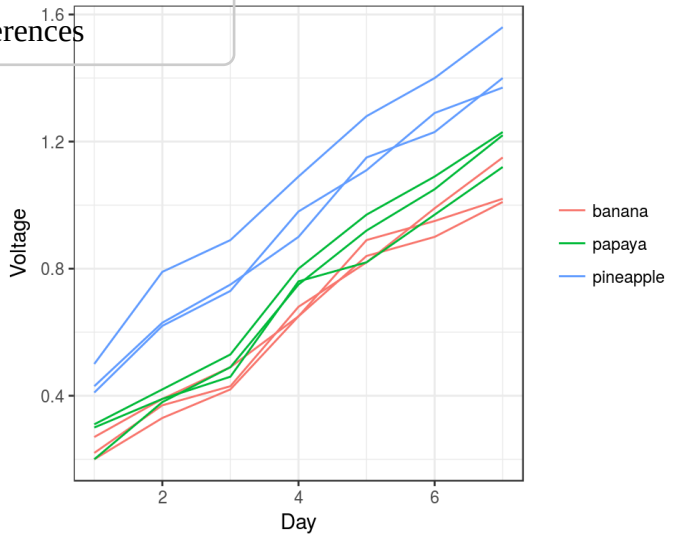
CODE



Here, it seems to be that the voltage outputs for banana and papaya do not seem to differ by much. On the other hand, the boxplots seem to suggest that voltage outputs for pineapple peelings are greater than that of the other two types of peelings.

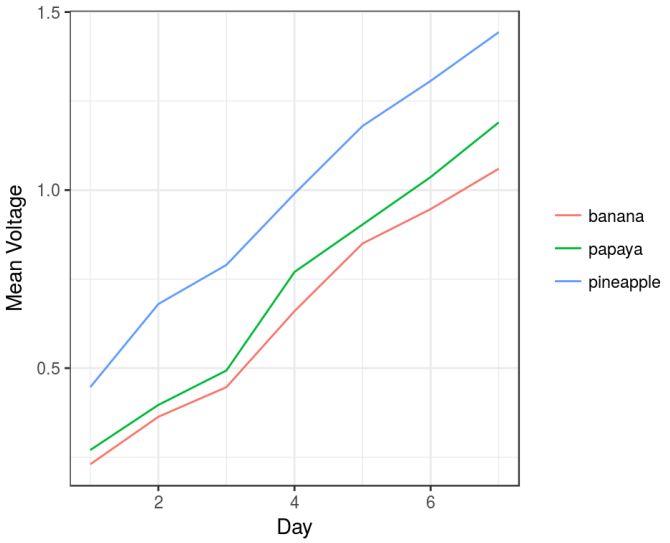
Get the plot of the voltages over 7 days for each type of fruit

CODE



Get the plot of the mean of the voltages over 7 days for each type of fruit

CODE



The line graphs suggest that voltage output increases directly proportional as the number of days.

Analysis of variance for repeated measures design

Your data set contains the following:

- Cell = blocking factor (random factor, since the effect each cell may contribute to the voltage is random)
- Day = within subject or repeated measures factor (fixed factor)
- fruit = the treatment, which is a between subject factor (fixed factor)
- Voltage = measured (dependent) variable

A one-way analysis of variance will not work here, nor will a randomized complete block design ANOVA will work, because the assumption of independence of observations is violated. This assumption is violated because observations are taken from the same cells over time, and are therefore dependent on the cells.

To model the variation in the voltage outputs, we can use both fixed effects and random effects factors in the data set. The model is called a **linear mixed effects model**. You can learn more about linear mixed effects model from the following links:

- http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf (http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf)
- https://wiki.bcs.rochester.edu/HlpLab/StatsCourses?action=AttachFile&do=get&target=Groningen11.pdf (https://wiki.bcs.rochester.edu/HlpLab/StatsCourses?action=AttachFile&do=get&

target=Groningen11.pdf)

- https://arxiv.org/ftp/arxiv/papers/1308/1308.5499.pdf (https://arxiv.org/ftp/arxiv/papers/1308/1308.5499.pdf)
- http://www.theanalysisfactor.com/repeated-measures-approaches/ (http://www.theanalysisfactor.com/repeated-measures-approaches/)

Based on your study, what we want to model are the following: * There might be interactions between time (Day) and treatment (fruit), that is, we may allow the effect of Day to vary between type of fruit peeling (fruit) * We will allow change over time to differ across participants (i.e., to explicitly model differences in voltage changes among cells over time). Because of this, we will also allow the effect of Day to be *random*.

With these considerations, our model can be written as:

```
Voltage ~ Day + fruit + Day:fruit + (Day|Cell)
```

The analysis will be carried in R (R Core Team, 2017), using the `lmerTest` (Kuznetsova, Bruun Brockhoff, & Haubo Bojesen Christensen, 2016) package for modeling linear mixed effects. Post-hoc analysis will be conducted with the help of the R package `multcomp` (Hothorn, Bretz, & Westfall, 2008).

CODE

```
## Analysis of Variance Table of type III
with  Satterthwaite
## approximation for degrees of freedom
##           Sum Sq Mean Sq NumDF DenDF F.v
alue  Pr(>F)
## fruit      0.064   0.032     2     6
16  0.0037
## Day        2.608   2.608     1     6
1334 2.8e-08
## fruit:Day  0.008   0.004     2     6
2  0.2178
```

From the linear mixed effects model ANOVA table, we can see that the interaction of the type of fruit peeling and time does not affect the voltage output ($F(2, 6) = 2, p = 0.2178$). We can therefore interpret the effects of the type of fruit peelings and Day individually without worrying about their possible interactions. Here, we are primarily interested with the effect of type of fruit peelings. We can see that the voltage outputs differ between types of fruit peelings ($F(2, 6) = 16, p == 0.0037$) at the 0.05 significance level.

To see which of the types of fruit peelings differ, we conduct a post-hoc analysis using a paired-t test with Bonferroni-Holm with multiple-comparisons adjustment using the `multcomp` package.

CODE

```
##
##   Simultaneous Tests for General Linear
Hypotheses
##
## Multiple Comparisons of Means: Tukey Con
trasts
##
##
## Fit: lme4::lmer(formula = Voltage ~ frui
t * Day + (Day | Cell), data = data)
##
## Linear Hypotheses:
##                          Estimate Std. Er
ror z value Pr(>|z|)
## papaya - banana == 0      0.0162     0.0
477    0.34     0.73
## pineapple - banana == 0   0.2438     0.0
477    5.11  9.7e-07
## pineapple - papaya == 0   0.2276     0.0
477    4.77  2.8e-06
## (Adjusted p values reported -- BH method
)
```

There is no difference in voltage outputs between papaya and banana peelings ($p = 0.73$). On the other hand, the differences in voltage outputs between pineapple and banana and pineapple and papaya are very extreme that they could not be due to chance alone ($p < 0.0001$ in both cases).

Addendum

Your objectives do not include the effect of the time on the voltage output. However, the ANOVA table suggests that the number of days affect the voltage output significantly ($F(1, 6) = 1334, p < 0.0001$). If you want to include the effect of the days in your objective, then we can see the coefficient due to Day.

CODE

```
## Linear mixed model fit by REML t-tests u
se Satterthwaite approximations
##     to degrees of freedom [lmerMod]
## Formula: Voltage ~ fruit * Day + (Day |
Cell)
##    Data: data
##
## REML criterion at convergence: -153.1
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -1.904 -0.469  0.002  0.613  2.357
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
Corr
##  Cell     (Intercept) 2.02e-03 0.04493
##           Day         9.53e-05 0.00976
-0.27
##  Residual             1.95e-03 0.04421
## Number of obs: 63, groups:  Cell, 9
##
## Fixed effects:
##                    Estimate Std. Error
df t value Pr(>|t|)
## (Intercept)         0.07095    0.03374 6
.00000    2.10   0.0801
## fruitpapaya         0.01619    0.04771 6
.00000    0.34   0.7459
## fruitpineapple      0.24381    0.04771 6
.00000    5.11   0.0022
## Day                 0.14500    0.00742 6
.00000   19.54  1.2e-06
## fruitpapaya:Day     0.01393    0.01049 6
.00000    1.33   0.2326
## fruitpineapple:Day  0.02048    0.01049 6
.00000    1.95   0.0989
##
## Correlation of Fixed Effects:
##             (Intr) frtppy frtpnp Day
frtpp:D
```

```
## fruitpapaya -0.707
## fruitpinppl -0.707  0.500
## Day         -0.530  0.375  0.375
## fruitpy:Dy   0.375 -0.530 -0.265 -0.707
## frtpnppl:Dy  0.375 -0.265 -0.530 -0.707
## 0.500
```

From this output, you can see that a day contributes about 0.145V (± 0.007) of voltage. However, your data has very few observations to determine when you will have achieved the maximum voltage output.

Furthermore, your sample is still small in order to say that the differences we have measured in this study are definitive. (You would need about 200 observations per type of peeling in order to achieve a power of 0.8 at .05 significance level.) This study is, therefore, more of an initial investigation type.

Reminders about the tables

- Re-write the outputz of ANOVA and the post-hoc analysis as nice-looking tables that are properly captioned.
- If in doubt, ask your adviser about the formatting.

References

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2016). *LmerTest: Tests in linear mixed effects models*. Retrieved from https://CRAN.R-project.org/package=lmerTest (https://CRAN.R-project.org/package=lmerTest)

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for

Statistical Computing. Retrieved from https://www.R-project.org/ (https://www.R-project.org/)