MT5763_220013309

Erna Kuginyte 220013309

2022-11-19

Abstract

This report summarises the statistical modelling and analysis findings related to the efficacy of a new baldness drug study for the Luxuriant company. The study's goal is to see if the new baldness drug Luxuriant has an effect that is greater than the placebo effect, if the new drug is more effective than existing treatments, and if the age of the population has any bearing on the effect. The standard statistical methods such as linear regression model fitting, two-sided two-sample Student's t-test, The Wilcoxon Rank Sum Test, Analysis of Variance (Anova) with Tukey HSD, and Kruskal-Wallis test are used. The report summarises that Luxuriate indeed is an effective treatment when compared to placebo, however, it does not produce better results than the products that already exist in the market. Age factor presented as not important.

Introduction

The aim of the study is to see if the new baldness drug Luxuriant has an effect that is greater than the placebo effect, whether the new drug is more effective than existing treatments, and if the age of the population has any bearing on the effect. The study was conducted for the statistical professionals in the "Luxuriant" company. The data used in the study was provided by the "Luxuriant" company; it is not extensive, but it does include information on two other baldness treatment products as well as the placebo effect. The "Baldy" data set includes: hair length post the new drug "Luxuriant" is considered in the same trial with both a placebo and two existing anti-baldness treatments "BaldBeGone" and "HairyGoodness",as well as the age of each participant. The patients, all of whom were more or less totally bald, were randomly allocated to the groups, shaved and then had any hair growth measured after one month of treatment. The standard statistical methods such as linear regression model fitting, two-sided two-sample Student's t-test and the non-parametric equivalent The Wilcoxon Rank Sum test, Analysis of Variance (Anova) with Tukey HSD and the non-parametric equivalent Kruskal Wallis test are used.

Methods

Analysis is done using SAS OnDemand for Academics.

Data wrangling is simple since the data did not contain any missing data, the measured hair length is converted from inches to millimeters. The original data frame is rearranged to have three columns containing Age, Drug and Length_mm values.

Exploratory analysis includes data normality test using Wilk-Shapiro test, a boxplot of hair growth in mm by treatment, boxplot of age values by treatment, scatterplot of hair growth in mm from age, grouped by treatment.

To check whether "Luxuriant" has an effect above Placebo, Student's T-Test is used. The Student's t-test is somewhat robust against not normally distributed data, especially since this study has enough samples for each treatment group. To be safe, an equivalent non-parametric Wilcoxon Rank Sum test is used. To find whether "Luxuriant" is more effective than the existing treatments on the market, Analysis of Variance (Anova) with Tukey HSD and the non-parametric equivalent Kruskal Wallis test are used. Age relevancy is checked by fitting a linear regression model.

Results

Table 1: First rows of wrangled data, example

Length_mm	Age	Drug
0.6628575	30	Luxuriant
0.5702790	31	Luxuriant
0.8727847	33	Luxuriant
0.4791224	33	Luxuriant
2.9013424	36	BaldBeGone

Confidence level refers to the percentage of probability, or certainty, that the confidence interval would contain the true population parameter when you draw a random sample many times.

Conclusions

The way the data has been conducted is slightly odd, and because natural hair growth rates differ between people, such a study might benefit from comparing the same group of people before and after "Luxuriant," "BaldBeGone," "HairyGoodness," and placebo treatments. Another important factor to point out is that since the sample size is not vast, it might not capture the true effect in the population. Other factors, such as hair type and normal hair growth rate prior to any treatment, may play a significant role. The study, overall, might not capture the true effect of the new drug, or even the true difference between the treatments, solely because of the way the data has been gathered. Examining a new group of people with suggested improvements is highly recommended. From the data that was available, the report summarises that Luxuriate indeed is an effective treatment when compared to placebo, however, it does not produce better results than the products that already exist in the market. The linear regression model revealed that age was not important.

References

https://www.stat.purdue.edu/~tqin/system101/method/method_wilcoxon_rank_sum_sas.htm

https://www.stat.purdue.edu/~tqin/system101/method/method_kruskal_wallis_sas.htm

Appendices

Code

Data Wrangling

```
/* Import Baldy data set */
filename refile '/home/u62740852/bald/baldy.csv';
proc import datafile = refile
    dbms = csv
    out = df;
    getnames = yes;
run;

/* Check contents of data */
proc contents data = df;
run;

/* Convert inch values to mm */
data df;
    set df;
    Luxuriant_mm = Luxuriant * 25.4;
    Placebo_mm = Placebo * 25.4;
    BaldBeGone_mm = BaldBeGone * 25.4;
    HairyGoodness_mm = HairyGoodness * 25.4;

/* Remove data columns with inch values */
    drop Luxuriant Placebo BaldBeGone HairyGoodness;
run;

/* Luxuriant table */
data df_l;
    set df (rename = (Luxuriant_mm = Length_mm AgeLuxuriant = Age));
    Drug = 'Luxuriant';
    drop AgePlacebo AgeBaldBeGone AgeHairyGoodness
         Placebo_mm BaldBeGone_mm HairyGoodness_mm;
run;

/* Placebo table */
data df_p;
    set df (rename = (Placebo_mm = Length_mm AgePlacebo = Age));
    Drug = 'Placebo';
    drop AgeLuxuriant AgeBaldBeGone AgeHairyGoodness
         Luxuriant_mm BaldBeGone_mm HairyGoodness_mm;
run;

/* BaldBeGone table */
data df_b;
    set df (rename = (BaldBeGone_mm = Length_mm AgeBaldBeGone = Age));
    Drug = 'BaldBeGone';
    drop AgeLuxuriant AgePlacebo AgeHairyGoodness
         Luxuriant_mm Placebo_mm HairyGoodness_mm;
run;

/* HairyGoodness table */
```

```
data df_h;
    set df (rename = (HairyGoodness_mm = Length_mm AgeHairyGoodness = Age));
    Drug = 'HairyGoodness';
    drop AgeLuxuriant AgePlacebo AgeBaldBeGone
         Luxuriant_mm Placebo_mm BaldBeGone_mm;
run;

/* Merge all drug tables together */
data df_1;
    set df_l
    df_p
    df_b
    df_h;
run;

/* Merge Luxuriant with Placebo to then run a t.test and Wilcoxon test */
data df_2;
    set df_l
    df_p;
run;

/* Check contents of data */
proc contents data = df_1;
run;
```

Normality Checks

```
/* Normality checks */
/* Plot histograms of each distribution to check normality visually, Lenght_mm */
ods graphics / reset width = 6.4in height = 4.8in imagemap;
proc sort data = df_1 out = _HistogramTaskData;
    by Drug;
run;
proc sgplot data = _HistogramTaskData;
    by Drug;
    title "Normality check, Hair Length";
    histogram Length_mm;
    density Length_mm;
    density Length_mm / type = kernel;
    keylegend / location = inside position = topright across = 1 noborder;
    yaxis grid;
run;
ods graphics / reset;
proc datasets library = WORK noprint;
    delete _HistogramTaskData;
run;

/* Plot histograms for normality check, Age */
ods graphics / reset width = 6.4in height = 4.8in imagemap;
proc sort data = df_1 out = _HistogramTaskData;
    by Drug;
run;
proc sgplot data = _HistogramTaskData;
    by Drug;
    title "Normality check, Age";
```

```
    histogram Age;
    density Age;
    density Age / type = kernel;
    keylegend / location = inside position = topright across = 1 noborder;
run;
ods graphics / reset;
proc datasets library = WORK noprint;
    delete _HistogramTaskData;
    run;

/* Perform an official Shapiro-Wilk test to test normality of data */
proc univariate data = df normal;
run;
```

Exploratory Analysis

```
/* Plot hair growth by drug. */
proc sgplot data = df_1;
    vbox Length_mm / category = Drug fillattrs = (color = cxCCCCFF)
               outlierattrs = (color = cx800080 size = 5pt);
    title 'Hair Growth by Treatment';
    xaxis label = 'Treatment' valueattrs = (style = italic)
               grid gridattrs = (color = cxC0C0C0);
    yaxis label = 'Hair Growth in mm' grid gridattrs = (color = cxC0C0C0);
run;

/* Boxplot age by drug */
proc sgplot data = df_1;
    vbox Age / category = Drug fillattrs = (color = cxCCCCFF)
               outlierattrs = (color = cx800080 size = 5pt);
    title 'Age by Treatment';
    xaxis label = 'Treatment' valueattrs = (style = italic)
               grid gridattrs = (color = cxC0C0C0);
    yaxis label = 'Age' grid gridattrs = (color = cxC0C0C0);
run;

/* Scatterplot length from age */
proc sgplot data = df_1;
    scatter x = Age y = Length_mm / group = Drug;
    title 'Hair Growth from Age';
    xaxis label = 'Age' grid gridattrs = (color = cxC0C0C0);
    yaxis label = 'Hair Growth in mm' grid gridattrs = (color = cxC0C0C0);
run;

/* Luxuriant scatter plots with regression and CIs added */
proc sgplot data = df_l;
  title "Luxuriant. Hair Growth from Age";
  reg x = Age y = Length_mm / CLM CLI;
  scatter x = Age y = Length_mm / transparency = 0.5 markerattrs = (symbol = circlefilled);
run;
title;

/* Placebo scatter plots with regression and CIs added */
proc sgplot data = df_p;
  title "Placebo. Hair Growth from Age";
```

```
  reg x = Age y = Length_mm / CLM CLI;
  scatter x = Age y = Length_mm / transparency = 0.5 markerattrs = (symbol = circlefilled);
run;
title;

/* BoldBeGone scatter plots with regression and CIs added */
proc sgplot data = df_b;
  title "BoldBeGone. Hair Growth from Age";
  reg x = Age y = Length_mm / CLM CLI;
  scatter x = Age y = Length_mm / transparency = 0.5 markerattrs = (symbol = circlefilled);
run;
title;

/* HairyGoodness scatter plots with regression and CIs added */
proc sgplot data = df_h;
  title "HairyGoodness. Hair Growth from Age";
  reg x = Age y = Length_mm / CLM CLI;
  scatter x = Age y = Length_mm / transparency = 0.5 markerattrs = (symbol = circlefilled);
run;
title;
```

Linear Model, Statistical Tests

```
/* Check if age is relevant, run a linear regression model */
proc reg data = df_1;
    model Length_mm = Age;
run;

/* Run two-sample two-tailed Student t-test for Luxuriant and Placebo treatments,
alpha = 0.05, null hypothesis H0 = mean1 - mean2 = 0, H1 = mean1 - mean2 != 0
Not normally distributed data, need to run a non-parametric test */
proc ttest data = df_2 sides = 2 alpha = 0.05 h0 = 0;
    title "Two sample two-sided t-test to compare Luxuriant and Placebo treatments";
    class Drug;
    var Length_mm;
run;

/* Wilcoxon Rank Sum test */
proc NPAR1WAY data = df_2 wilcoxon;
    title "Nonparametric test to compare Luxuriant and Placebo treatments";
    class Drug;
    var Length_mm;
    exact wilcoxon;
run;

/* Run Tukey HSD test, Luxuriant and HairyGoodness have similarity
Not normally distributed data, need to run a non-parametric test */
proc glm data = df_1;
    class Drug;
    model Length_mm = Age Drug;
    lsmeans Drug / adjust = tukey;
run;

/* Kruskal-Wallis non parametric test */
proc npar1way data = df_1;
```

```
    class Drug;
    var Length_mm;
run;
```