Decision Analysis Report
Analysis of Location and Type of Bambola that Yield the Largest Profit

Minjiao Yang

Contents

1 Summary

The main purpose of this report is to find the best combination of Location and Variety in planting Bambola seeds, such that Bambola Corporation could make most profit from 10000. This report consists of 4 major sections.Introduction gives some background information for this experiment and raises several questions providing a direction for a further study. Data Summary part study the distribution of Color; Germination rate under various Location and Variety; transformation for two response Height and Total. Then, the Analysis section. Firstly, we try to predict log(Height) by Variety, Location and Color. Started with the Full Model including 3 variables and all interactions, through a sequential elimination procedure, the final model only includes location and variety. RMSE is 0.419, all assumptions are met and (FL:B) yields a biggest height. Secondly, we try to predict total profit with our model, and conclude that (FL, D) is most profitable combination.

The last is the conclusion part. It provides the answer to several questions in which we are interested in the early analysis. We conclude that: the distribution of COLOR in our data-set does not have significant difference with White 57%, Pink 36%, the 7% Red; Germination rate is not significantly different among varieties but significantly different between FL and LA; (FL, C) produced significantly more Red canes than the others; (FL, B) is the most favorable condition for canes to get height; (FL, D) is most profitable combination. So I recommend plant type D bambola in Florida will yield a total profit with 95%CI in (1,012,687, 1,556,313) dollars.

2 Introduction

The Bambola Corporation aims to find the most profitable plan to plant 10,00 Bambola seeds. Specifically, the client wants to know which combination of Location and Variety is the best. Four variety of Bambolas were planted in the past, where 256 seeds of each (1024 seeds in all), half of each variety in the FL plot and half in the LA plot. Each planted seed may or may not germinate, but most of them germinated and grew into cane, with the distribution believed to be about 57% White, 36% Pink, and 7% Red. The value of each cane is determined by both color and height. If the cane was 1 meter or longer in height, the price is 0.20/cm, 0.30/cm, and 0.50/cm (in dollars) for White, Pink, and Red canes, respectively. If the cane was less than 1 meter in height, then was chopped into 10 cm slices that were made into whistles to be sold with the price 1, 2, and 4 (dollars) per whistle for White, Pink, and Red whistles, respectively. Whistles should be at least 10 cm, so cane slice shorter than 10 cm is valueless. Except for the most profitable Variety and Location combination of Bambola planting, client is also interested in the following questions:

1. For the overall experiment (not controlling for Variety or Location) is there any evidence that the [W:P:R] color distribution is significantly different than stated before?

2. Non-germination rate for each one of the varieties. Is the probability of non-germination significantly different between varieties A and D?

3. Is there a difference between the germination rates in the two locations?

4. Is there one Variety and/or Location that produces significantly more Red canes than the others?

5. Find a model for cane Height (given that the plant germinates) as a function of Variety, Location, Color and/or their interactions. And find the most favorable condition.

6. For each combination of Variety and Location, compute the expected total value for a crop of 10,000 seeds.

3 Data Summary

Name	Explanation
SEQ	Plant Sequence ID Number(1-1024)
LOC	Location(FL or LA)
VAR	Variety (A, B, C, D)
LIVE	Indicator of Germination of Seed (1=Yes, 0=No)
CLR	Color of Cane [White(W), Pink(P), Red(R)]
HT	Height (in cm.) of Cane
WHIST	Value (in US dollar) of Cane Whistles
POLES	Value (in US dollar) of Cane Pole
TOTAL	Total Value (in US dollar) of Cane

Table 1: Explanation of variables in Bambola experiment

Most of the planted seeds germinated(LIVE=1) and grew into canes. If seeds never germinated, then no information about cane (color, height, value) can be obtained. If seeds grow into canes then its height and color are recorded, with longer cane that is 1 meter or more are made into finishing pole, and price for different color are also different. White cane can be valued to \$ 0.20/cm,White cane can be valued to \$ 0.30/cm and red for \$ 0.50/cm. If the cane was less than 1 meter then it was chopped into 10cm slices(length short than 10cm were discarded for no value)that were made into whistles to be sold at tourist shops, with the price per whistle being \$1, \$2 and \$4 for white, pink and red whistles. The Bambola Corporation want find out what combination of Variety and location would be expected to yield the most profit, and they will plant 10,000 seeds of this variety at this location and would like to see how much money this would yield. If varieties are all about equally good, then they will use Variety (easy to obtain), IF locations are about same, they would prefer Florida(more accessible).

4 Exploratory Data Analysis

4.1 Estimate Germination

4.1.1 Germination by Location

To estimate germination for each variety or location, I preform the chi-square independence test and logistic regression for germination or non-germination vs Variety or location. and result are shown in following tables. By performing the chi-square independence test, with null hypothesis be all variety with same non-germination rate, and I obtain the p-value for chi-square test is 0.4193, which implies I cannot reject my null hypothesis. So there's no significantly difference for non-germination rate among four varieties. Which also tells us there's no significantly difference for germination rate among four varieties.

Variety	Germinated	Non-germinated	Total	P_{germ}	$P_{non-germ}$
A	228	28	256	89.1%	10.9%
B	231	25	256	90.2%	9.8%
C	225	31	256	87.9%	12.1%
D	236	20	256	92.2%	7.8%
Total	920	104	1024	89.8%	10.2%

Table 2: Germination for each variety

4.1.2 Germination by Location

Next, I use same method to estimate the germination rate for two locations.

Location	germinated	non-germinated	Total	P_{germ}	$P_{non-germ}$
FL	441	71	512	86.1%	13.9%
LA	479	33	512	93.6%	6.4%
Total	920	104	1024	89.8%	10.2%

Table 3: Germination for each location

By performing the chi-square independence test, with null hypothesis be all locations with same germination rate, and I obtain the p-value for chi-square test is 0.0001293, which implies I can reject my null hypothesis. So there's significantly difference for germination rate among two locations.

For further investigation, I run the logistic model to see the probability of germination for each location, and check if it consistent with my previous result.

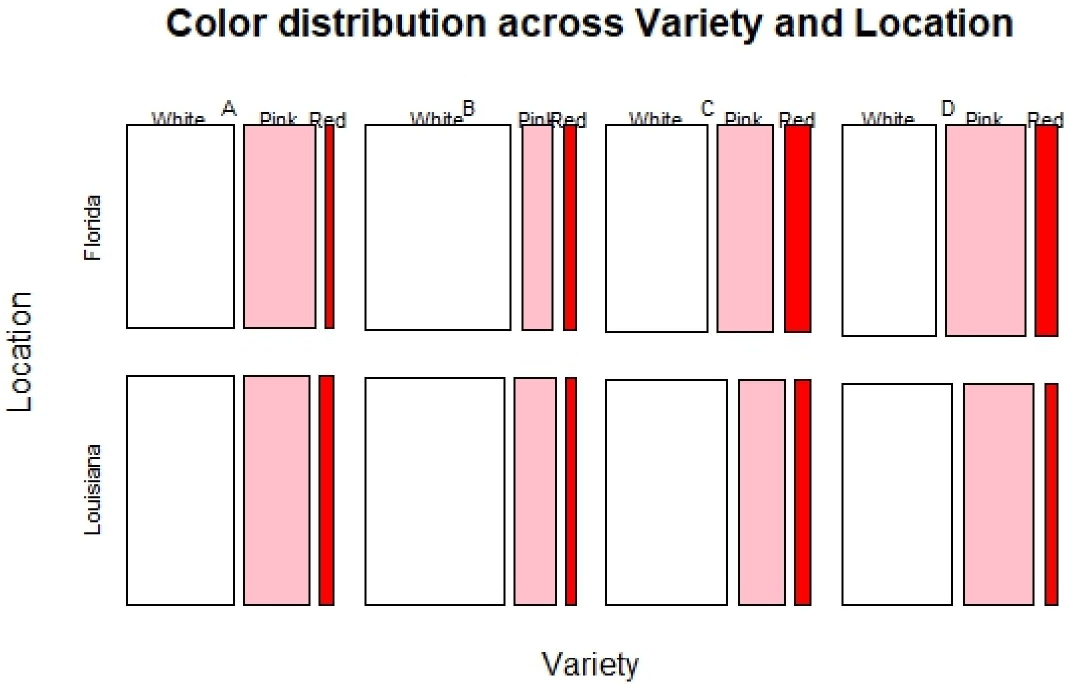
4.2 Investigate Color Distribution

People believed that the distribution of color of cane is (White,Pink,Red)=(57% ,36% ,7%), I run the chi-square test of goodness of fit to see if our collected data followed this distribution.

Color	Believed Distribution	Real Distribution
White	57%	62%
Pink	36%	30%
Red	7%	8%

Table 4: Color Distribution of Cane

Since our p-value is small, we reject our null hypothesis that color distribution are same as people believed and conclude that Color distribution is significantly different from people believed. Furthermore, by plotting the Mosic plot, it is easier to see which variety and/or location produces more Red canes.



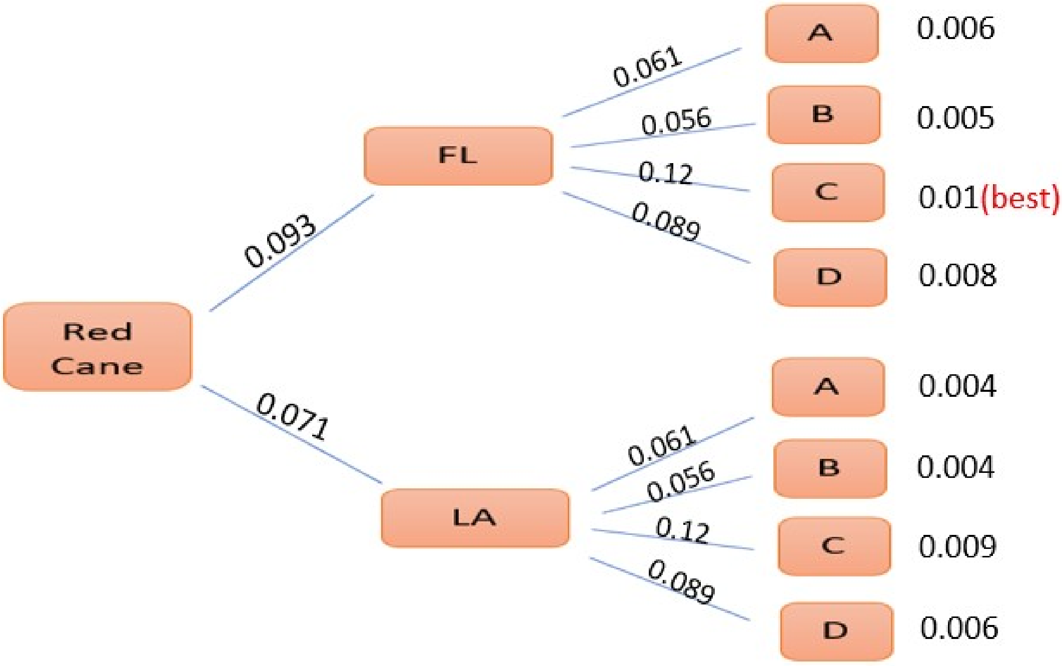
From picture, Florida and variety C is obviously produced more red. But we need further investigation with chi-square independence and logistic regression tests.By performing the chi-square independence test, with null hypothesis be all locations with same red appearance rate, and I obtain the p-value for chi-square test is 0.2726(table 17 in Appendix), which implies I cannot reject my null hypothesis. So there's no significantly difference for red appearance rate among two locations. From table 6 and 8, It's clearly that Variety C is significantly produce more red canes than any other. So, we have Variety C in Florida produces red cane the most(see table 7).

Location	Red	Total	P_{red}
FL	41	441	9.3%
LA	34	479	7.1%

Table 5: Red Cane for each location

Variety	Red	Total	P_{red}
A	14	228	6.1%
B	13	231	5.6%
C	27	225	12%
D	21	236	8.9%

Table 6: Red Cane for each variety



Variety	Location	
	FL	LA
A	4.7%	7.4%
B	6.4%	5.0%
C	14.8%	9.4%
D	11.2%	6.7%

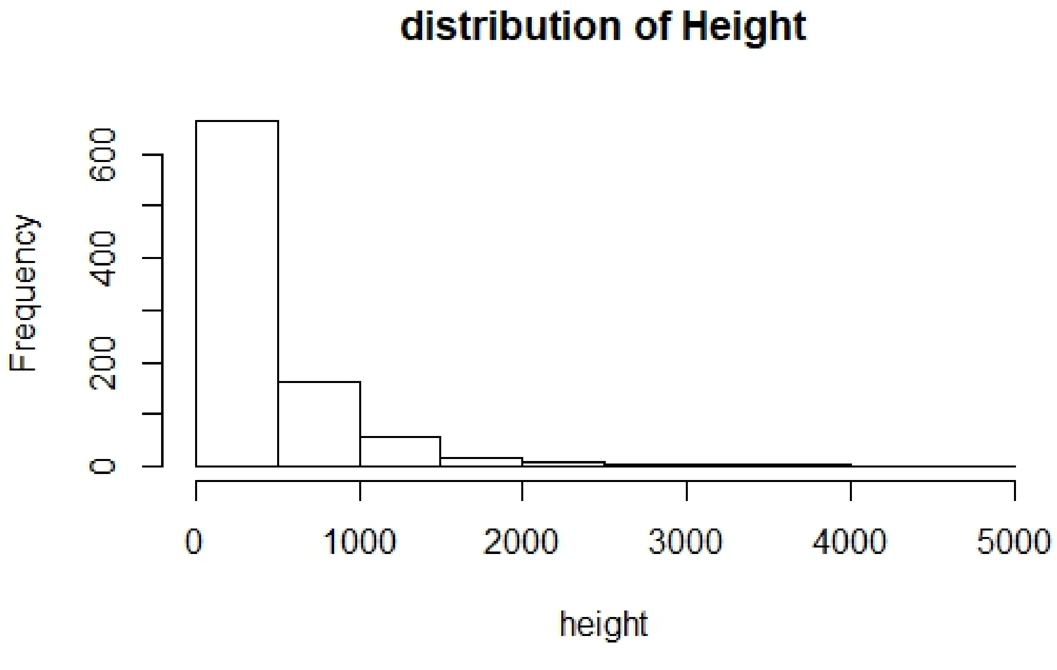
Table 7: Red Cane for each variety in each location

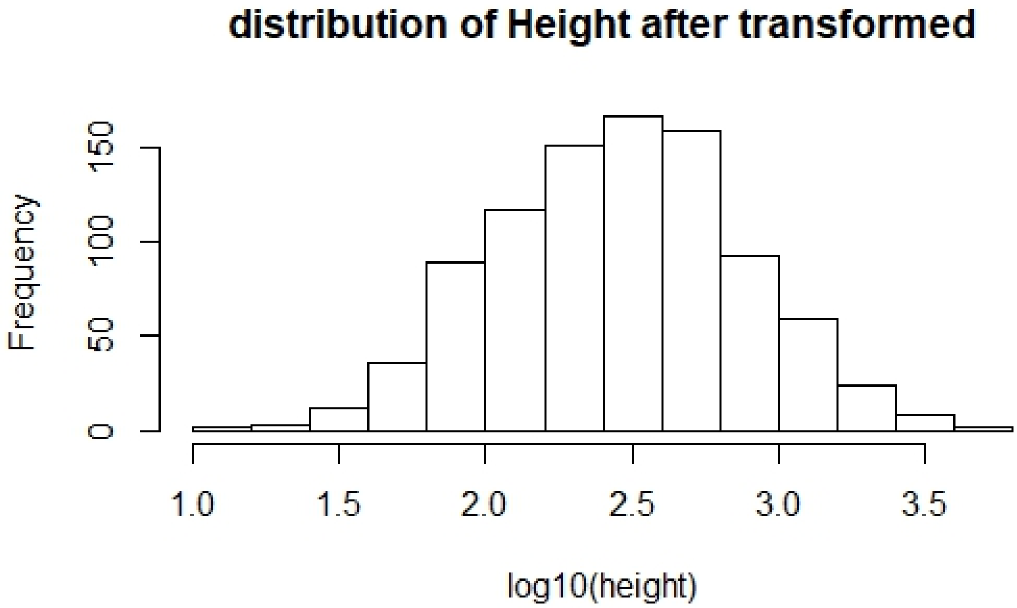
Coefficients	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-2.72692	0.27586	-9.885	<2e-16
Var(B)	-0.09263	0.39700	-0.233	0.8155
Var(C)	0.73449	0.34379	2.136	0.0326
Var(D)	0.40080	0.35829	1.119	0.2633

Table 8: Red Cane for each variety

5 Model for Cane Height

By looking at the distribution of height of cane, we need to first do transformation of it since it is not normally distributed, which is skewed with longer tail.





After transformed my height and made it normally distributed. then I found the model for height with two main factors variety and location compare it with my mean table for each combination.

	FL	LA
A	2.49	2.45
B	2.53	2.49
C	2.50	2.31
D	2.52	2.34

Table 9: mean for each variety-location combination

In table 9, Florida B has largest mean value, to see if it's estimated height yield the largest, I use following model to run the regression test:
$log10(Height) = 2.5315 - 0.1150 \times I(LA)0.0371 \times I(B) - 0.0721 \times I(C) - 0.0451 \times I(D)$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.5315	0.0314	80.63	0.0000
factor(LOC)LA	-0.1150	0.0277	-4.16	0.0000
factor(VAR)B	0.0371	0.0391	0.95	0.3429
factor(VAR)C	-0.0721	0.0394	-1.83	0.0674
factor(VAR)D	-0.0451	0.0389	-1.16	0.2468

Table 10: model for cane height

From table 10, with FL and A as baseline, FL is significantly better than LA, B is better than A,C and D. which is consistent with our mean table for each combination.Therefore, Variety B in

Florida is the most favorable.

5.1 The Variety-Location that yield the Most Profit

	FL	LA
A	128	128
B	128	128
C	128	128
D	128	128

Table 11: size for total value

	FL	LA
A	102.49	121.80
B	108.15	106.99
C	110.70	70.29
D	128.45	76.66

Table 12: mean for total value

	FL	LA
A	157.77	174.54
B	150.76	130.55
C	152.28	90.13
D	155.88	80.04

Table 13: sd for total value

FL & D seems to be the highest mean, but $SE = SD/\sqrt{128}$ is from $ 11-15 are large, most of them are not significant different, mean for LAC and LAD is small and Se are small, they are significant worsen than other. Since (FL,D) yield the greatest profit, when plant 10,000 seeds, we will get the expected total value be $E(T) = 10,000 \times \$128.45 = \$1,284,500$
$SE = \frac{\$155.88}{\sqrt{128}} = \$13.78$
$SE(T) = \$13.78 \times 10,000 = \$13,7800$
$SD(T) = 100 \times \$155.88 = \$15,588$
$SD(T) = \sqrt{(\$15,588)^2 + (\$13,7800)^2} = \$138,680$
And 95% prediction Interval for total value of (FL, D) is $(E(T) - 1.96 \times SD(T), E(T) + 1.96 \times SD(T)) = (\$1,012,687, \$1,556,313)$ So Florida and Variety D expect to yield total value between $ 1,012,687 and $ 1,556,313.

6 Conclusion

Based on our analysis above, we can conclude that:

1. The distribution of COLOR in our data-set has not significant difference with White 57%, Pink 36%, the 7% Red.
2. Non-germination rate not significantly different between varieties A and D.
3. Non-germination rates are significantly different between location LA and FL.
4. The combination (FL, C) produced significantly more Red canes than the others.
5. The combination (FL, B) is the most favorable condition for canes to get height.
6. (FL, D) is most profitable combination. So I recommend plant type D bambola in Florida will yield a total profit with 95%CI in (1,012,687, 1,556,313) dollars.

7 Appendix

1. Germination for each variety

$X^2 = 2.8254$	df = 3	p-value = 0.4193

Table 14: Pearson's Chi-squared test

2. Germination for each Location

$X^2 = 14.652$	df = 1	p-value = 0.0001293

Table 15: Pearson's Chi-squared test

3. Color Distribution of Cane

$X^2 = 13.272$	df = 2	p-value = 0.001312

Table 16: Pearson's Chi-squared test for given probability

4. Red Cane for each location

$X^2 = 1.2036$	df = 1	p-value = 0.2726

Table 17: Pearson's Chi-squared test

5. Red Cane for each Variety

$X^2 = 7.8231$	df = 3	p-value = 0.04981

Table 18: Pearson's Chi-squared test

6. Red Cane for each variety in each location

$X^2 = 12.376$	df = 7	p-value = 0.08884

Table 19: Pearson's Chi-squared test