Midterm Project

James Young

July 11, 2019

The Problem:

A farmer wants to know which seeding rate gives the best yield with reasonable statistical power. The farmer supplies preliminary data from four fields, each field counting as one experimental unit. Given the data, which will be described below, I will determine how many replicates are necessary to run a statistically powered experiment in the future, which would aim to determine optimal seeding rate.

The Data:

I have four csv files, one each for the four fields I have been given data from. Each file has six columns (Sample, Easting, Northing, AppliedRate, ControlRate, and Yield). Preliminary exploratory data analysis shows each field has a different number of samples with with the smallest being 6718 samples and the largest being 12654, almost twice the size. Also the field with the smallest sample size had 8 rows with NA for yield. I chose to drop the rows with NA, rather than impute a yield, as it is generally bad practice to impute.

Methods:

I will individually aggregate each of the four csv files by ControlRate which will give the mean of all the columns by different levels of ControlRate. I will then merge the four aggregated files into a new file keeping only ControlRates 23000, 24000,…,29000 which I justify with an explanation in the code below. This results in each ControlRate having 3 to 4 datapoints along with the corresponding yield. I turned these datapoints into averages and standard deviations which were then used to make summary statistics and to calculate effect size and replicates needed.

Effect size looks at how meaningful a difference between two treatments is when given the mean and standard deviation for the two treatments (larger effect size is more meaningful). A small effect size is > 0.2, a moderate effect size > 0.5, and a large effect size > 0.8. An effect size of 1 is the equivalent of a differnce of 1 standard deviation. The equation is given below as “cohen.d”.

Required replicates, which is the primary output of this work, are calculated when given the means and standard deviation of two treatments and also an alpha and beta value. I will hold beta constant at 0.2 and have three levels of alpha 0.1, 0.05, and 0.01 (liberal, moderate, and conservative, respectively) to make different estimates on how many replicates are needed. A conservative estimate is more stringent and makes a statistically powered experiment more likely. The equation is given below as “required.replicates”.

The final function I make below is “integerize”, which I use to turn the estimated required replicates into integers. Any estimate below 1 is turned into 1 and any estimate above 1 is rounded to the nearest integer.

Code and Intermediate Results:

cohen.d = function(M1, S1, M2, S2){  
 SD = (((S1\*\*2+S2\*\*2)/2)\*\*(1/2))  
 d = ((M1 - M2)) / SD  
 return(d)  
}  
  
  
required.replicates <- function (m1, s1, m2, s2, alpha, beta) {  
 CV <- (((s1\*\*2+s2\*\*2)/2)^(1/2))/((m1+m2)/2)  
 Diff <- (m1-m2)/((m1+m2)/2)  
 sdpooled <- ((((s1\*\*2 + s2\*\*2))/2)\*\*(1/2))  
 2 \* ((CV/Diff)\*\*2) \* (((1-(alpha/2))+(1-beta)))\*\*2}  
  
integerize <- function(a){  
 a = trunc(a)  
 is.na(a) <- !a  
 a[is.na(a)] <- 1  
 a  
}

Here, I load the data, check dimensions, remove NA’s, recheck dimensions, and gather summary statistics.

#Load data from four fields  
fieldA <- read.csv("C://data/fieldA.csv")  
fieldB <- read.csv("C://data/fieldB.csv")  
fieldC <- read.csv("C://data/fieldC.csv")  
fieldD <- read.csv("C://data/fieldD.csv")  
  
dim(fieldA)

## [1] 6718 6

dim(fieldB)

## [1] 9321 6

dim(fieldC)

## [1] 10404 6

dim(fieldD)

## [1] 12654 6

fieldA <- na.omit(fieldA)  
fieldB <- na.omit(fieldB)  
fieldC <- na.omit(fieldC)  
fieldD <- na.omit(fieldD)  
  
dim(fieldA)

## [1] 6710 6

dim(fieldB)

## [1] 9321 6

dim(fieldC)

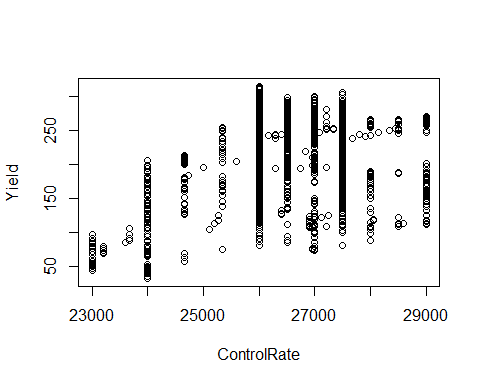
## [1] 10404 6

dim(fieldD)

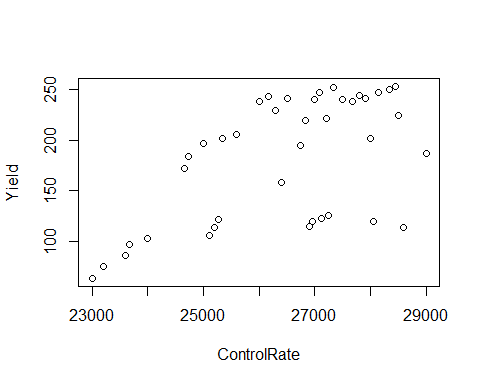
## [1] 12654 6

Below I compare what the unaggregated and aggregated data look like when plotting Yield against ControlRate for field A. IT looks roughly like a quadratic equation in both graphs.

plot(data = fieldA, Yield ~ ControlRate)



aggfieldA <- aggregate(fieldA, by=fieldA["ControlRate"], FUN=mean, na.rm=TRUE)  
plot(data = aggfieldA, Yield ~ ControlRate)



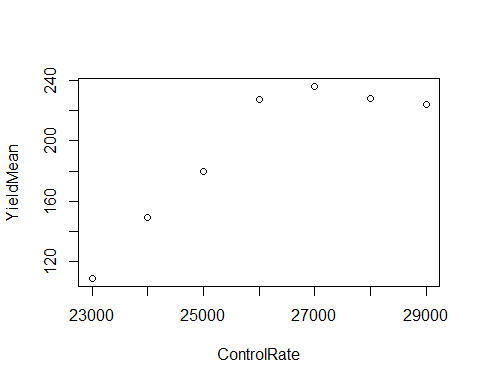
Here I aggregate the rest of the fields in the same manner as fieldA.

aggfieldB <- aggregate(fieldB, by=fieldB["ControlRate"], FUN=mean, na.rm=TRUE)  
aggfieldC <- aggregate(fieldC, by=fieldC["ControlRate"], FUN=mean, na.rm=TRUE)  
aggfieldD <- aggregate(fieldD, by=fieldD["ControlRate"], FUN=mean, na.rm=TRUE)

Now I merge the fields into one dataset. I only take data that has ControlRate 23000, 24000,…,29000. I chose this option becuase binning other values into these levels seems like it would be unfair to the high and low bounds of ControlRate. For example, if 23000-23500 were binned to 23000, there is not a compensatory 29000-29500 range on the upper bound, since the highest ControlRate is 29000. Therefore, I chose to only accept values that had the exact ControlRate we were given in previous work. This still results in 3 to 4 data points per ControlRate.

I then plot the data to visualize the trend between ControlRate and Yield.

merged <- rbind(aggfieldA, aggfieldB, aggfieldC, aggfieldD)  
  
  
merged <- subset(merged, ControlRate ==23000 | ControlRate == 24000 | ControlRate == 25000 | ControlRate == 26000 | ControlRate == 27000 | ControlRate == 28000 | ControlRate == 29000)  
  
merged2 = aggregate(merged$Yield, by=merged["ControlRate"], FUN=mean, na.rm=TRUE)  
merged2$YieldMean <- merged2$x  
merged2$x <- NULL  
  
merged3 = aggregate(merged$Yield, by=merged["ControlRate"], FUN=sd, na.rm=TRUE)  
merged2$YieldSD <- merged3$x  
plot(data=merged2, YieldMean ~ ControlRate)



summary(merged2)

## ControlRate YieldMean YieldSD   
## Min. :23000 Min. :108.6 Min. : 6.016   
## 1st Qu.:24500 1st Qu.:164.4 1st Qu.:13.680   
## Median :26000 Median :224.3 Median :24.002   
## Mean :26000 Mean :193.4 Mean :24.133   
## 3rd Qu.:27500 3rd Qu.:227.8 3rd Qu.:34.116   
## Max. :29000 Max. :236.3 Max. :43.323

It looks like a ControlRate of 2700 produces the highest yield, therefore I will use 2700 as the standard which I calculate effect size and required replicates against.

Here, I calculate liberal, moderate, and conservative estimates of required reps using the previosly discussed “required.replicates” function. The difference in estimates is a result of varying the alpha value to 0.1, 0.05, and 0.01 as previously discussed.

Moderate\_reqreps = required.replicates(236.3188, 6.015556, merged2$YieldMean, merged2$YieldSD, 0.05, 0.2)  
Conservative\_reqreps = required.replicates(236.3188, 6.015556, merged2$YieldMean, merged2$YieldSD, 0.01, 0.2)  
Liberal\_reqreps = required.replicates(236.3188, 6.015556, merged2$YieldMean, merged2$YieldSD, 0.1, 0.2)  
EffectSize = cohen.d(236.3188, 6.015556, merged2$YieldMean, merged2$YieldSD)

Here I round calculated required replicates to integers. If a calculation is below 1 it is rounded up to 1 and if a calculation is above 1 it is rounded to the nearest integer.

Liberal = integerize(Liberal\_reqreps)  
Moderate = integerize(Moderate\_reqreps)  
Conservative = integerize(Conservative\_reqreps)

Here, I bind the ControlRates, EffectSize, and Liberal, Moderate, and Conservative estimates of required replicates.

ControlRate <- merged2$ControlRate  
Final <- cbind(ControlRate, EffectSize, Liberal, Moderate, Conservative)  
Final <- as.data.frame(Final)  
#Drop 27000 control rate since that is what we are comparing against  
Final <- Final[-c(5),]  
Final

## ControlRate EffectSize Liberal Moderate Conservative  
## 1 23000 4.1302690 1 1 1  
## 2 24000 3.4493032 1 1 1  
## 3 25000 3.9687904 1 1 1  
## 4 26000 1.2312541 4 4 4  
## 6 28000 0.4756635 27 27 28  
## 7 29000 0.5054396 23 24 25

Results and Conclusions:

Looking at the final data table, ControlRate of 27000 has large effect size vs. all of the lower ControlRates and a small to moderate effect size versus all of the higher ControlRates. If you want to compare 27000 to 26000 an estimated 4 replicates are needed, regardless of if the estimate is conservative (more stringent) or not. If you want to compare 27000 to 28000 an estimated 27 or 28 replicates are needed, depending on how conservative you want to be. The most striking feature is that 26000 looks as visually close to 27000 as 28000 or 29000 do, but has a larger calculated effect size, possibly making this the most interesting test and comparison. After 27000 the yield drops with a small to moderate effect size. This may mean, upon further testing, that 27000 is the optimal ControlRate for yield return when analyzed with the methods used here.