Experimental Power Calculation

Statistical Power Calculation

2/13/2020

Pairs: Sherry Zuo, Yishuang Song

1. You are a marketing manager at Nordsaksingdale's, the retailer that ran the experiment in Assignment #1. You are now planning a second experiment and are choosing between several options. In this larger experiment, you expect to be able to show ads to **500,000** people (total across treatment and control). You need to decide how many of these to allocate to the control group (with a PSA ad) vs. the treatment group (with a Nordsaksingdale's ad featuring Givenchy hand bags). You also need to choose how much budget to allocate to the ad campaign. To guide your choice, you run several statistical power calculations.

One important ingredient in the power calculation is the standard deviation of the sales. You make two assumptions:
1) the standard deviation of sales is the same as in the previous experiment; and 2) you expect the standard deviation of sales in the new experiment to be the same in both the treatment and control groups.
What is the standard deviation of sales (not past_sales) in the treatment group among exposed users who saw the campaign in "AdFX-Class-TermRow_count.csv" (from Assignment 1)? Use this number for your calculations below.

```
#load data
AdFx<-read_csv("AdFX-BA860-SectionB-W20-4004106-rows.csv")
```

```
## Parsed with column specification:
## cols(
##   Treatment = col_double(),
##   saw_ads = col_double(),
##   sales = col_double(),
##   past_sales = col_double(),
##   gender = col_character()
## )
```

```
Treatment_group<-AdFx%>%filter(Treatment==1)
Treatment_exposed<-Treatment_group%>%
  filter(saw_ads==1)
sd_sales<-sd(Treatment_exposed$sales)
sd_sales
```

```
## [1] 4.960137
```

Answer: The standard deviation of sales in the treatment group among exposed users who saw the campaign in "AdFX-Class-TermRow_count.csv" is **4.960137** dollars.

2. Continuing from Q1, suppose that you have the budget to spend $**0.01** per person on advertising. As your 'reasonable signal' or benchmark AdFX lift, you assume that the campaign will break even (0 profit).

a. If your profitability margin on sales is **50%**, what is the 'reasonable signal' for AdFX lift that you are assuming? (2 points)

```
margin=0.5
cost=0.01
revenue=cost#since the campaign will break even
signal=revenue/margin
signal
```

```
## [1] 0.02
```

Answer: The reasonable signal (benchmark AdFX lift) is **0.02** dollars.
reasonable signal = $0.02
standard deviation of sales = $4.960137
total sample size = 0.5 million
d = signal/ standard deviation

b. For a 95% confidence interval, calculate the statistical power to detect a successful campaign for the following three potential experimental designs

i). 20% of users are assigned to the control group

```
pwr.t2n.test(d=0.02/4.960137, n1=0.2*5*10^5, n2=(1-0.2)*5*10^5, sig.level = .05)
```

```
##
##      t test power calculation
##
##              n1 = 1e+05
##              n2 = 4e+05
##               d = 0.004032147
##       sig.level = 0.05
##           power = 0.207216
##     alternative = two.sided
```

ii). 30% of users are assigned to the control group

```
pwr.t2n.test(d=0.02/4.960137, n1=0.3*5*10^5, n2=(1-0.3)*5*10^5, sig.level = .05)
```

```
##
##      t test power calculation
##
##              n1 = 150000
##              n2 = 350000
##               d = 0.004032147
##       sig.level = 0.05
##           power = 0.2572931
##     alternative = two.sided
```

iii). 80% of users are assigned to the control group

```
pwr.t2n.test(d=0.02/4.960137, n1=0.8*5*10^5, n2=(1-0.8)*5*10^5, sig.level = .05)
```

```
##
##      t test power calculation
##
##              n1 = 4e+05
##              n2 = 1e+05
##               d = 0.004032147
##       sig.level = 0.05
##           power = 0.207216
##     alternative = two.sided
```

Which design is best according to this criterion? (9 points)
Answer: The statistical power of 20% Control is 0.207216, the statistical power of 30% Control is 0.2572931, The statistical power of 80% Control is 0.207216. Since 0.2572931 is larger than 0.207216, the design **ii (30% Control)** is best according to this criterion.

c. What treatment assignment (% assigned to the control group) maximizes statistical power? (Hint: use solver, or play with the spreadsheet) (5 points)

```
power <- function(x){
  pwr.t2n.test(d = 0.02/4.960137, n1 = x*5*10^5, n2 = (1-x)*5*10^5, sig.level = 0.05)$power
}
optimize(f = power, upper = 1, lower = 0, maximum = T)
```

```
## $maximum
## [1] 0.5
##
## $objective
## [1] 0.2968919
```

Answer: **50% Control assignment** maximizes statistical power, which is 0.2968919

3. Continuing from Q1 & Q2, Now suppose instead that you have a fixed ad budget of $2,000 to spend on your own ads. You ignore the cost of control ads because your partner the publisher is providing these for free.

a. Calculate the average ad spend per person for the three experimental designs above. (6 points)

```
#2 i
avg1<-2000/((1-0.2)*5*10^5)
avg1
```

```
## [1] 0.005
```

```
#2 ii
avg2<-2000/((1-0.3)*5*10^5)
avg2
```

```
## [1] 0.005714286
```

```
#2 iii
avg3<-2000/((1-0.8)*5*10^5)
avg3
```

```
## [1] 0.02
```

Answer: The average ad spend per person for the experimental design 2bi(20% Control) is **0.005** dollars; The average ad spend per person for the experimental design 2bii(30% Control) is **0.005714286** dollars; The average ad spend per person for the experimental design 2biii(80% Control) is **0.02** dollars.

b. Maintaining your assumption that the campaign will break-even for each experimental design, what is the 'reasonable signal' for each of the three possible experimental designs above? (6 points)

```
rs1=avg1/margin
rs1
```

```
## [1] 0.01
```

```
rs2=avg2/margin
rs2
```

```
## [1] 0.01142857
```

```
rs3=avg3/margin
rs3
```

```
## [1] 0.04
```

Answer: The reasonable signal for the experimental design 2bi(20% Control) is **0.01**; The reasonable signal for the experimental design 2bii(30% Control) is **0.01142857**; The reasonable signal for the experimental design 2biii(80% Control) is **0.04**.

c. For a 95% confidence interval, what is the statistical power now for each of the three possible experimental designs? Which is highest now? (9 points)

```
pwr.t2n.test(d=rs1/4.960137, n1=0.2*5*10^5, n2=(1-0.2)*5*10^5, sig.level = .05)
```

```
##
##      t test power calculation
##
##              n1 = 1e+05
##              n2 = 4e+05
##               d = 0.002016073
##       sig.level = 0.05
##           power = 0.08800489
##     alternative = two.sided
```

```
pwr.t2n.test(d=rs2/4.960137, n1=0.3*5*10^5, n2=(1-0.3)*5*10^5, sig.level = .05)
```

```
##
##      t test power calculation
##
##              n1 = 150000
##              n2 = 350000
##               d = 0.002304084
##       sig.level = 0.05
##           power = 0.1158958
##     alternative = two.sided
```

```
pwr.t2n.test(d=rs3/4.960137, n1=0.8*5*10^5, n2=(1-0.8)*5*10^5, sig.level = .05)
```

```
##
##      t test power calculation
##
##              n1 = 4e+05
##              n2 = 1e+05
##               d = 0.008064293
##       sig.level = 0.05
##           power = 0.6258901
##     alternative = two.sided
```

Answer: The statistical power now for the experimental design 2bi(20% Control) is **0.08800489**; The the statistical power now for the experimental design 2bii(30% Control) is **0.1158958**; The the statistical power now for the experimental design 2biii(80% Control) is **0.6258901**. The design **iii** is highest now which is the **80% Control**.

d. What changes your answer between Q2b) and Q3c)? Which is the best of the six options in terms of statistical power? (5 points)

Answer: The cost of control ads is changed, which causes the change of **effect size(d)**. The best is **Q3c) iii** since it has the highest statistical power which is 0.6258901 since it's larger than 0.2572931, so the best is **the 80% control, 20% treatment when cost of control ads is free**.

4. Continuing on Q1 to Q3, you have been thinking about the possibility your ads may wear out so that their effectiveness decreases as you increase the average ad spend. You revisit your previous assumption and instead assume the following for the 3 possible experimental designs:
Design Cost Per Person 'Reasonable' Signal

```
rs4=avg1*2
rs4
```

```
## [1] 0.01
```

```
rs5=avg2*1.9
rs5
```

```
## [1] 0.01085714
```

```
rs6=avg3*1.2
rs6
```

```
## [1] 0.024
```

a. What is the 'reasonable' signal now? Fill out the above table. (6 points)

Answer:
20% Control: Reasonable Signal: 2X cost = ***0.01***
30% Control: Reasonable Signal: 1.9X cost = ***0.01085714***
80% Control: Reasonable Signal: 1.2X cost = ***0.024***

b. Under your revised assumption, what is the statistical power of each option? Which is best now? (9 points)

```
pwr.t2n.test(d=rs4/4.960137, n1=0.2*5*10^5, n2=(1-0.2)*5*10^5, sig.level = .05)
```

```
##
##      t test power calculation
##
##              n1 = 1e+05
##              n2 = 4e+05
##               d = 0.002016073
##       sig.level = 0.05
##           power = 0.08800489
##     alternative = two.sided
```

```
pwr.t2n.test(d=rs5/4.960137, n1=0.3*5*10^5, n2=(1-0.3)*5*10^5, sig.level = .05)
```

```
##
##      t test power calculation
##
##              n1 = 150000
##              n2 = 350000
##               d = 0.00218888
##       sig.level = 0.05
##           power = 0.1093254
##     alternative = two.sided
```

```
pwr.t2n.test(d=rs6/4.960137, n1=0.8*5*10^5, n2=(1-0.8)*5*10^5, sig.level = .05)
```

```
##
##      t test power calculation
##
##              n1 = 4e+05
##              n2 = 1e+05
##               d = 0.004838576
##       sig.level = 0.05
##           power = 0.2775592
##     alternative = two.sided
```

Answer: The the statistical power now for the experimental design 2bi(20% Control) is **0.08800489**; The the statistical power now for the experimental design 2bii(30% Control) is **0.1093254**; The the statistical power now for the experimental design 2biii(80% Control) is **0.2775592**. The design **iii** is highest now which with **80% Control**

5. One vexing problem for Nordsaksingdale's has been to measure the effects of its paid search advertising on in-store sales. This is challenging because search platforms do not offer database match campaigns and they do not allow advertisers to experiment at the user level.
   Your marketing team has come up with an intriguing solution to this problem. Since search platforms let you target by designated marketing area (DMA), you can experiment by advertising in some markets and 'going dark' (turning off ads) in others. Then, you will compare in-store sales in the treatment and control markets. For the test, the marketing team has selected 60 DMAs that each has a single Nordsaksingdale's retail locations and split them equally into treatment and control markets. You are excited about the opportunity to learn about this medium's effectiveness through an experiment. Nonetheless, you first want to evaluate the experiment's statistical power before committing resources.

a. You budget $40,000 weekly for the paid search campaign. For your power calculation, you expect the search ads to generate a lift equivalent to 3 X the ad cost. What is the expected ad lift for the average store (the 'reasonable' signal)? (2 points)

```
lift_store=3*(4*10^4)/30
lift_store
```

```
## [1] 4000
```

Answer: The expected ad lift for the average store is **4000** dollars.

b. Across the 60 locations, the stores have the same average weekly sales of 200,000 dollars with standard deviation 30,000 dollars. Given this and using a 90% confidence interval as your standard, calculate the statistical power for this experiment if you run it for a single week. (4 points)

```
pwr.t2n.test(d=lift_store/30000, n1=0.5*60, n2=0.5*60, sig.level = .1)
```

```
##
##      t test power calculation
##
##              n1 = 30
##              n2 = 30
##               d = 0.1333333
##       sig.level = 0.1
##           power = 0.143871
##     alternative = two.sided
```

Answer: The statistical power for this experiment is **0.143871** if I run it for a single week.

c. At current spending levels, how many weeks would you need to run this experiment before the statistical power of the experiment surpasses 50%? (4 points)

```
store_power=0
i=0
while(store_power<0.5) {
  i=i+1
  test_power <- pwr.t2n.test(d=lift_store/30000, n1=0.5*60*i, n2=0.5*60*i, sig.level = .1)
  store_power <- test_power$power
  print(store_power)
}
```

```
## [1] 0.143871
## [1] 0.1879865
## [1] 0.2310856
## [1] 0.2730193
## [1] 0.3136656
## [1] 0.3529297
## [1] 0.3907411
## [1] 0.4270507
## [1] 0.4618278
## [1] 0.4950579
## [1] 0.5267398
```

```
weeks <- i-1
weeks
```

```
## [1] 10
```

Answer: The weeks = 10, at current spending levels, we need **10** weeks to run this experiment before the statistical power of the experiment surpasses 50%.

6. After many years at your Nordsaksingdale's marketing job, you make a career move to become the CMO of IAMS pet food (congratulations!). You advertise online, but can't link your ads to sales data directly because more than 90% of your sales occur in-store via grocery and pet stores. Nonetheless, measurement is important to you and you are evaluating a proposal by DynamicLogic, a market research firm. DynamicLogic will run an experiment combined with surveys to determine how your ads affect consumer perception of your brand (as we discussed in class). From past surveys, you know that your baseline brand favorability is 4.1 out of 5 (standard deviation 0.68) and your intent to purchase measure is 1.6 out of 5 (s.d. 1.23). Suppose you expect the effect of seeing mobile ads to improve brand favorability by 2% and intention to purchase by 4%. You are planning an experiment that shows the IAMS ads and a survey to 300,000 users as well as a control ad and a survey to 100,000 users. When pushed, DynamicLogic explains that they expect about 0.2% of people who see the survey to fill it out.

a. Given the response rates, how many surveys do you expect to collect in each of the treatment and control groups? (2 points)

```
300000*0.002#Treatment
```

```
## [1] 600
```

```
100000*0.002#Control
```

```
## [1] 200
```

Answer: I expect to collect **600 surveys in treatment group** and **200 surveys in control group**.
b. DynamicLogic survey proposal is expensive and you can only pay them for this one ad campaign. What is the likelihood that the 95% confidence intervals on your AdFX estimates exclude 0 for each survey measure? (4 points)

```
##ask for statistical power
lift_brand_favorability=4.1*0.02
lift_brand_favorability
```

```
## [1] 0.082
```

```
lift_intention_purchase=1.6*0.04
lift_intention_purchase
```

```
## [1] 0.064
```

```
pwr.t2n.test(d=lift_brand_favorability/0.68, n1=600, n2=200, sig.level = .05)
```

```
##
##      t test power calculation
##
##              n1 = 600
##              n2 = 200
##               d = 0.1205882
##       sig.level = 0.05
##           power = 0.31419
##     alternative = two.sided
```

```
pwr.t2n.test(d=lift_intention_purchase/1.23, n1=600, n2=200, sig.level = .05)
```

```
##
##      t test power calculation
##
##              n1 = 600
##              n2 = 200
##               d = 0.05203252
##       sig.level = 0.05
##           power = 0.09754985
##     alternative = two.sided
```

Answer: The likelihood that the 95% confidence intervals on your AdFX estimates exclude 0 for baseline brand favorability is **0.31419**, for intent to purchase measure is **0.09754985**

c. In a one-paragraph summary for a coworker, explain your decision on whether to use DynamicLogic's services. Be sure to discuss the strengths and limitations of survey analysis and explain implications of your power calculation for the experiment. (6 points)

Answer: The strengths of survey analysis is that it is low-costing, and data information can be gathered easily to do further analysis. While the limitations are also obvious, even though survey can give insights and represent people's opinions about ads, it cannot fully explain the influence on purchase rate. Additionally, people may actually purchase nothing even if they choose intent to purchase. The statistical power of baseline brand favorability is 0.31419, which is lower then 50% and not high enough, and the statistical power of intent to purchase measure is 0.09754985, which is very low, so we do not have enough confidence on AdFX estimates exclude 0. Because of the low click-through rates, the number of sample is also too small. Therefore, these numbers do not prove the validity of this experiment and we **do not recommend** using DynamicLogic's services.

d. Suppose that 0.5% of users who see the treatment ad fill out the survey, but 0.1% of users who see the control ad fill the survey. How would this affect your interpretation of the experimental results? (6 points) *Answer:* The difference in completion rates between the treatment and control groups is troubling. Experiments are predicated on the treatment group and control group being the 'same' due to the experimental randomization. In the case of a survey experiment, we further require that those who choose to take the survey are the 'same' in both groups for the results to be valid. Here, the difference in response rates raises a red flag that the two groups of respondents are different ahead of time. We can speculate that, for instance, the treatment ad may have garnered more attention than the control ad such that more and different kinds of people responded to the treatment ad survey