

Answer Sketch for Homework 6

431 Staff and Professor Love

'Due 2018-11-09, version 2018-10-25

Contents

Initial R Setup	1
1 Question 1 (25 points)	2
2 Question 2 (10 points)	2
3 Question 3 (10 points)	3
4 Question 4 (10 points)	4
Setting up Questions 5-7	4
5 Question 5 (10 points)	5
6 Question 6 (10 points)	5
7 Question 7 (10 points)	6

Initial R Setup

Here's the R setup we used.

```
knitr::opts_chunk$set(comment=NA)
options(width = 70)

library(Epi); library(magrittr); library(tidyverse)
```

Later, we'll also use the `mosaic` package (for `favstats`).

We'll use the `twobytwo` function here, from the `Love-boost.R` script on our Data and Code page.

```
`twobytwo` <-
function(a,b,c,d, namer1 = "Row1", namer2 = "Row2",
         namec1 = "Col1", namec2 = "Col2",
         conf.level = 0.95)
  # build 2 by 2 table and run Epi library's twoby2 command to summarize
  # from the row-by-row counts in a cross-tab
  # upper left cell is a, upper right is b, lower left is c, lower right is d
  # names are then given in order down the rows then across the columns
  # use standard epidemiological format - outcomes in columns, treatments in rows
  {
    require(Epi)
    .Table <- matrix(c(a, b, c, d), 2, 2, byrow=T,
                    dimnames=list(c(namer1, namer2),
                                   c(namec1, namec2)))
    twoby2(.Table, alpha = 1 - conf.level)
  }
```

Or we could have simply sourced in the `Love-boost.R` file.

1 Question 1 (25 points)

Silver writes (in several places prior to Chapter 12, but especially there) that the goal of any predictive model is to capture as much signal as possible and as little noise as possible. What does this mean to you in your scientific and other endeavours, going forward? Give a specific example. An answer of 150 - 250 words is what we're looking for.

We don't write sketches for essays.

2 Question 2 (10 points)

Suppose you plan to study whether surgery can prolong life among men suffering from prostate cancer, which typically develops and spreads very slowly. Men diagnosed with prostate cancer will be randomly assigned to either undergo surgery or not. Suppose you believe that approximately 10% of men diagnosed with prostate cancer but do not have surgery will eventually die from prostate cancer, and you want to do the study using a sample size that will retain at least 80% power to detect a drop down to 5%, using a two-sided approach with a 95% confidence level.

What is the smallest number of men (including both the surgery and non-surgery groups) that you will need to enroll in this new study to meet these specifications, using a balanced design? Provide the details of your calculation, and also provide a complete sentence or two interpreting the meaning of the results in context.

We use the `power.prop.test` function to determine this sample size, with the probability without surgery at $p_1 = 0.10$, and the probability after surgery at $p_2 = 0.05$, at a 5% significance level with 80% power.

```
power.prop.test(p1 = 0.10, p2 = 0.05, sig.level=0.05, power = 0.80)
```

Two-sample comparison of proportions power calculation

```
      n = 434.432
      p1 = 0.1
      p2 = 0.05
sig.level = 0.05
  power = 0.8
alternative = two.sided
```

NOTE: `n` is number in *each* group

We will require at least 435 men in each of the two arms of the study - meaning that we need a total of **870 men**.

3 Question 3 (10 points)

*In 2003, the **New England Journal of Medicine** published results of a Scandinavian study of this issue. In that study, among 347 men randomly assigned to surgery, 16 eventually died of prostate cancer, compared with 31 of the 348 men who did not have surgery.*

With 95% confidence, can you conclude that the odds of death due to prostate cancer is significantly greater for those who did NOT have surgery than it is for those who did have surgery? Provide appropriate details of your calculations, including a sentence or two interpreting the meaning of the results in context.

We'll build the two by two table with the rows describing surgical status (the exposure) and the columns describing mortality due to prostate cancer (the outcome). We'll put death due to prostate cancer and no surgery in the top left cell so as to obtain the odds ratio that is specified. I will go ahead and use the Bayesian augmentation here, but that's certainly not critical.

```
twobytwo(31+1, 348-31+1, 16+1, 347-16+1,  
         "No surgery", "Surgery", "PC Death", "No PC Death")
```

2 by 2 table analysis:

Outcome : PC Death

Comparing : No surgery vs. Surgery

	PC Death	No PC Death	P(PC Death)	95% conf. interval
No surgery	32	318	0.0914	0.0654 0.1264
Surgery	17	332	0.0487	0.0305 0.0769

	95% conf. interval
Relative Risk: 1.8770	1.0623 3.3165
Sample Odds Ratio: 1.9652	1.0699 3.6096
Conditional MLE Odds Ratio: 1.9634	1.0344 3.8497
Probability difference: 0.0427	0.0046 0.0819

Exact P-value: 0.0371

Asymptotic P-value: 0.0294

The sample odds ratio is 1.97 [2.02 without the Bayesian augmentation], and the 95% confidence interval for that odds ratio is (1.07, 3.61) [it would be (1.09, 3.77) without the augmentation] so there is a statistically significant difference in prostate cancer death rates, with higher odds of prostate cancer-related death associated with men who did not have the surgery. This is true because the 95% confidence interval for the appropriate odds ratio is greater than 1.

4 Question 4 (10 points)

In a 2014 follow-up report describing results at the end of 2012, a total of 63 men assigned to surgery and 99 men not assigned to surgery had died of prostate cancer. In a sentence or two, explain how an analysis of these new results compares to the conclusions you drew in Question 3.

We'll use the same approach as in Question 3.

```
twobytwo(99+1, 348-99+1, 63+1, 347-63+1,
         "No surgery", "Surgery", "PC Death", "No PC Death")
```

2 by 2 table analysis:

Outcome : PC Death

Comparing : No surgery vs. Surgery

	PC Death	No PC Death	P(PC Death)	95% conf. interval
No surgery	100	250	0.2857	0.2408 0.3353
Surgery	64	285	0.1834	0.1462 0.2275

		95% conf. interval
Relative Risk:	1.5580	1.1817 2.0543
Sample Odds Ratio:	1.7812	1.2468 2.5449
Conditional MLE Odds Ratio:	1.7797	1.2288 2.5907
Probability difference:	0.1023	0.0396 0.1641

Exact P-value: 0.0017

Asymptotic P-value: 0.0015

There is no substantial change in our conclusions. The sample odds ratio is now only 1.78 [1.79 without the Bayesian augmentation], but the 95% confidence interval for that odds ratio is (1.25, 2.55) [it would be (1.25, 2.57) without the augmentation] so there is still a statistically significant difference in prostate cancer death rates, with higher odds of prostate cancer-related death associated with men who did not have the surgery.

Setting up Questions 5-7

The `zocazo.csv` file shows the measurements of zocazolamine hydroxylase production (nmol 3H₂O formed / g per hour) for 13 women who smoked during pregnancy and 11 who did not.

```
zocazo <- read.csv("zocazo.csv") %>% tbl_df()
head(zocazo) # see first six observations
```

```
# A tibble: 6 x 3
```

	subject	production	group
	<int>	<dbl>	<fct>
1	1	3.56	smoker
2	2	0.56	non-smoker
3	3	10.1	smoker
4	4	0.36	non-smoker
5	5	16.5	smoker
6	6	3.36	smoker

5 Question 5 (10 points)

Develop a 99% confidence interval for the difference in the true means of zoxazolamine hydroxylase production in placentas from women who smoked as compared to those who did not, assuming that the distributions of production are approximately Normally distributed in each group.

These are independent samples of smokers and non-smokers. Assuming a Normal distribution in each group, we'd be inclined to use a two-sample t test.

```
t.test(production ~ group, data = zocazo,  
       conf.level = 0.99, var.equal = TRUE)
```

Two Sample t-test

```
data: production by group  
t = -3.238, df = 22, p-value = 0.003778  
alternative hypothesis: true difference in means is not equal to 0  
99 percent confidence interval:  
 -9.9418083 -0.6881218  
sample estimates:  
mean in group non-smoker      mean in group smoker  
      0.4327273              5.7476923
```

Here, we could either use a pooled t test, which yields a p value of 0.0038 or a Welch t test, which yields a p value of 0.0041, so it doesn't make an important difference.

Based on the pooled t test, our 99% confidence interval is (-9.94,-0.69).

This means that we are 99% confident that the difference in the true (population) means of zoxazolamine hydroxylase production in placentas from women who smoked minus those from women who did not smoke is between 0.69 and 9.94 nmol 3H₂O formed / g per hour.

6 Question 6 (10 points)

Suppose that in a new study, we assume a minimum clinically important effect 20% as large as was seen in the Kapitulnik study, and we assume a standard deviation of 1.5. If each individual measurement costs \$150 to obtain, how much money will be required to do the study with 99% confidence and 90% power?

The effect size (sample mean difference [smoker - non-smoker]) was 5.315 in the Kapitulnik study, so 20% of that is 1.063. The standard deviation is assumed to be 1.5, and we want to do the study with a 1% significance level, and 90% power.

```
mosaic::favstats(production ~ group, data = zocazo)
```

	group	min	Q1	median	Q3	max	mean	sd	n
1	non-smoker	0.18	0.36	0.42	0.53	0.68	0.4327273	0.1521244	11
2	smoker	0.60	1.37	3.56	9.00	16.50	5.7476923	5.4233556	13
	missing								
1		0							
2		0							

```
power.t.test(delta=0.2*(  
  zocazo %>% filter(group == "smoker") %$% mean(production) -  
  zocazo %>% filter(group == "non-smoker") %$% mean(production)),  
  sd = 1.5, power = 0.90, sig.level = 0.01)
```

Two-sample t test power calculation

```
n = 60.93847
delta = 1.062993
sd = 1.5
sig.level = 0.01
power = 0.9
alternative = two.sided
```

NOTE: n is number in *each* group

The required sample size is 61 measurements in each of the two groups, for a total of 122 measurements, each of which costs 150 dollars to obtain, so the total cost would be \$18300.

- Note that I used the `sprintf` command with the argument `%.0f` to get the total cost to print out without decimals or scientific notation.

7 Question 7 (10 points)

Suppose our maximum allowable budget is \$15,000 for the study. Comment on whether we can still do the study described in the previous question if we switched to a 95% confidence level.

So, again, we have a minimum clinically meaningful effect of size 1.063, with an assumed standard deviation of 1.5, and we want to do the study now with a 5% significance level, and 90% power. One way to assess whether our budget will finish the job would be to simply do the same sort of calculation to find the sample size required, as in Question 10.

```
power.t.test(delta=0.2*(
  zocazo %>% filter(group == "smoker") %$% mean(production) -
  zocazo %>% filter(group == "non-smoker") %$% mean(production)),
  sd = 1.5, power = 0.90, sig.level = 0.05)
```

Two-sample t test power calculation

```
n = 42.82897
delta = 1.062993
sd = 1.5
sig.level = 0.05
power = 0.9
alternative = two.sided
```

NOTE: n is number in *each* group

It turns out that we'd need 43 measurements in each group, or 86 total, and at \$150 per measurement, the total cost would be \$12900, which is less than our budget of \$15,000 for the study. So the answer is yes, we could do the study under these conditions.

Another way to see that this would be the case would be to recognize that with a budget of \$15,000, we can have up to 50 measurements in each group (thus 100 total, and a total cost of \$15,000.) So we could estimate the power we'd get under that circumstance, and see that it is 93.9%, which exceeds our requirement of 90%.

```
power.t.test(delta=0.2*(
  zocazo %>% filter(group == "smoker") %$% mean(production) -
```

```
zocazo %>% filter(group == "non-smoker") %>% mean(production),  
sd = 1.5, n = 50, sig.level = 0.05)
```

Two-sample t test power calculation

```
n = 50  
delta = 1.062993  
sd = 1.5  
sig.level = 0.05  
power = 0.9392386  
alternative = two.sided
```

NOTE: n is number in *each* group

As it turns out, we could still make this work up to about 97.5% confidence...

```
power.t.test(delta=0.2*(  
  zocazo %>% filter(group == "smoker") %>% mean(production) -  
  zocazo %>% filter(group == "non-smoker") %>% mean(production)),  
sd = 1.5, n = 50, sig.level = 0.025)
```

Two-sample t test power calculation

```
n = 50  
delta = 1.062993  
sd = 1.5  
sig.level = 0.025  
power = 0.8954896  
alternative = two.sided
```

NOTE: n is number in *each* group