Assignment 4 Answer Sketch

431 staff

due 2017-10-20 at noon

Contents

R Setup . . . 2

Question 1 (10 points) 2

Question 2 (10 points) 2

Question 3 (10 points) 2

Question 4 (20 points) 2

Question 5 (10 points) 3

Question 6 (10 points) 3

Question 7 (not graded - if you failed to do it, you lost 10 points, but otherwise, ignored by graders) 4

Question 7a. Outcome . . . 4
Question 7b. Exposure/Treatment Groups . . . 4
Question 7c. Paired or Independent Samples . . . 4
Question 7d. Random Sample? Representative? . . . 4
Question 7e. Significance Level? . . . 4
Question 7f. One-sided or Two-Sided Inference? . . . 5
Question 7g. Did pairing help (to reduce nuisance variation)? . . . 5
Question 7h. What does the distribution of sample paired differences tell us about which inferential procedure to use? . . . 5
Question 7i does not apply. These are paired samples. . . . 6
Question 7j. Did a statistically significant ($\alpha = .05$) change occur after treatment? Include in your response evidence supporting any assumptions you make. . . . 6

Question 8 (50 points) 8

Question 8a. Outcome (5 points) . . . 8
Question 8b. Exposure/Treatment Groups (5 points) . . . 8
Question 8c. Paired or Independent Samples (5 points) . . . 8
Question 8d. Random Sample? Representative? (5 points) . . . 9
Question 8e. Significance Level? (5 points) . . . 9
Question 8f. One-sided or Two-Sided Inference? (5 points) . . . 9
Question 8i. What does the distribution of each individual sample tell us about which inferential procedure to use? (10 points) . . . 9
Question 8j. On the basis of the most appropriate 95% CI, can you conclude that the outcome differs significantly between "control" and "histamine shock?" Justify your assumptions. (5 points) . . . 10
Setting up Questions 9-11 . . . 11

Question 9 (10 points) 12

Question 10 (10 points) 12

Question 11 (10 points) 13

R Setup

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

library(pander)
library(gridExtra)
library(tidyverse)

source("Love-boost.R")
```

We are grading questions 1-6 and 8-11, for a total of 150 points.

Question 1 (10 points)

Specify the URL where we can see the headline and news story describing the findings of the study. You should provide a complete reference, including the names of the author(s) of the news story, and its full title, and source.

Full credit for any working reference that included the specified information.

Question 2 (10 points)

Specify a URL where we can see at least the abstract of the complete study. Again, provide a complete reference to the study, too.

Full credit for any working reference that included the specified information.

Question 3 (10 points)

Describe your opinion (gut feeling) related to the conclusions of the study as summarized in the headline and news article, first in terms of a probability statement, and then calculate the appropriate odds. Motivate your internal prior probability, describing your relevant personal experiences or other factors that drove your gut feeling.

We want to see an accurate calculation given your probability, and at least a reasonable attempt at motivation for your initial probability.

Question 4 (20 points)

Evaluate the study in terms of the six specifications proposed by Leek when evaluating study support.

We want to see a clear, motivated conclusion about each of the six specifications, as well as direct quotes and evidence summaries to address the issues raised and justify conclusions.

The six specifications we are looking for are:

1. Was the study a clinical study in humans?

2. Was the outcome of the study something directly related to human health like longer life or less disease? Was the outcome something you care about, such as living longer or feeling better?
3. Was the study a randomized, controlled trial (RCT)?
4. Was it a large study - at least hundreds of patients?
5. Did the treatment have a major impact on the outcome?
6. Did predictions hold up in at least two separate groups of people?

Question 5 (10 points)

Incorporate the study support assessment into a Bayes' Rule calculation to obtain the final odds you should now be willing to give to the headline, and specify this value in terms of a probability statement, as well.

We wanted to see correct calculations of both the odds and probability in light of the prior probability established in Question 3 and the answers to the specifications described in Question 4.

Question 6 (10 points)

React to the final conclusion specified by this approach in a sentence or two. How does your subjective posterior probability that the headline is true match up with the formula's conclusions? Do you feel that the formulaic approach has yielded an appropriate conclusion for you in this case? Why or why not?

We wanted to see you specify your subjective feeling about what the probability *should* be, and then match that up with the result of the calculation.

Question 7 (not graded - if you failed to do it, you lost 10 points, but otherwise, ignored by graders)

For 10 diabetic adults treated with a special diet, the fasting blood sugar values (in mg/dl) before and after treatment were as shown below.

```
## Option 1 - create a hw4q7 data set in Excel, and import it into R
## using read.csv and tbl_df to make it into a tibble.
##
## Option 2 - build the tibble directly in R
hw4q7 <- tibble(
    person = LETTERS[1:10],
    before = c(340, 335, 220, 285, 320, 230, 190, 210, 295, 270),
    after = c(290, 315, 250, 280, 311, 213, 200, 208, 279, 258)
  )

pander(t(hw4q7)) # the t transposes rows and columns - not necessary
```

person	A	B	C	D	E	F	G	H	I	J
before	340	335	220	285	320	230	190	210	295	270
after	290	315	250	280	311	213	200	208	279	258

Question 7a. Outcome

The outcome is the before - after difference in fasting blood sugar (in mg/dl).

Question 7b. Exposure/Treatment Groups

The treatment is the special diet. The two groups are the patients before the diet and the same patients after the diet.

Question 7c. Paired or Independent Samples

These are paired samples. The data are paired by subject. Each subject provides a `before` and an `after` result.

Question 7d. Random Sample? Representative?

This isn't a random sample from the population of interest - which could be defined as adult patients with diabetes. The treatments were not assigned at random. Each patient provides a difference in blood sugar before and after the diet. It should be reasonable to treat this small sample of 10 observations as not clearly biased towards any particular end of the true population distribution, and so in that sense, we can think of it as representative.

Question 7e. Significance Level?

We are using a 5% significance level, or, equivalently, a 95% confidence level.

Question 7f. One-sided or Two-Sided Inference?

We are looking for a significant change in either a positive or negative direction, so that's a two-sided approach.

Question 7g. Did pairing help (to reduce nuisance variation)?

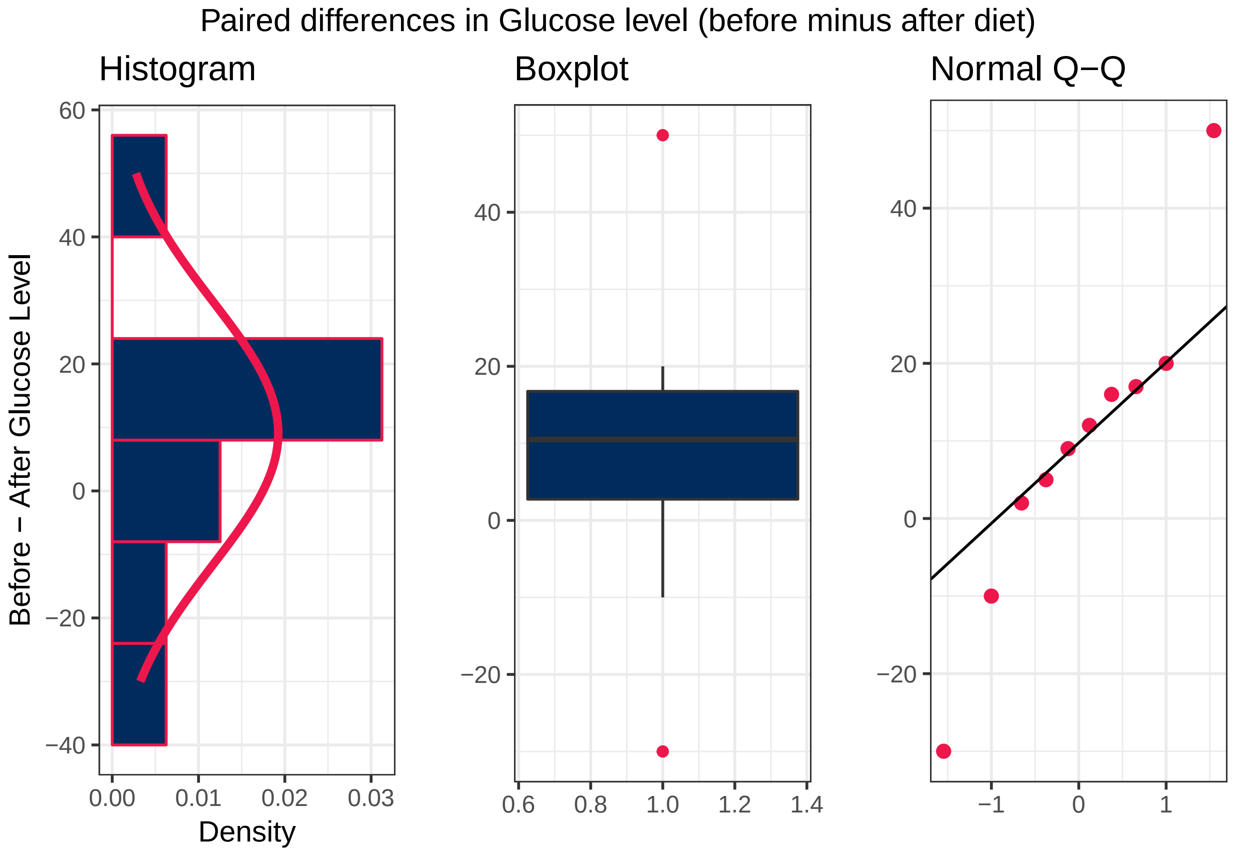
The correlation of the `before` and `after` glucose levels is 0.94, which is positive and strong, suggesting that there is a strong connection between the before and after scores of the subjects. Pairing definitely helped.

Question 7h. What does the distribution of sample paired differences tell us about which inferential procedure to use?

Here is our usual set of three plots describing the 10 paired (before - after) differences in glucose level, shown in the colors of the Cleveland Indians[1].

```r
hw4q7$diffs <- hw4q7$before - hw4q7$after

p1 <- ggplot(hw4q7, aes(x = diffs)) +
  geom_histogram(aes(y = ..density..), bins = fd_bins(hw4q7$diffs),
                 fill = "#002B5C", col = "#ED174C") +
  stat_function(fun = dnorm,
                args = list(mean = mean(hw4q7$diffs),
                            sd = sd(hw4q7$diffs)),
                lwd = 1.5, col = "#ED174C") +
  coord_flip() +
  theme_bw() +
  labs(title = "Histogram",
       x = "Before - After Glucose Level", y = "Density")

p2 <- ggplot(hw4q7, aes(x = 1, y = diffs)) +
  geom_boxplot(fill = "#002B5C", outlier.color = "#ED174C") +
  theme(axis.text.x = element_blank(),axis.ticks.x = element_blank()) +
  theme_bw() +
  labs(title = "Boxplot",
       y = "", x = "")

p3 <- ggplot(hw4q7, aes(sample = diffs)) +
  geom_qq(col = "#ED174C", size = 2) +
  geom_abline(intercept = qq_int(hw4q7$diffs),
              slope = qq_slope(hw4q7$diffs)) +
  theme_bw() +
  labs(title = "Normal Q-Q",
       y = "", x = "")

grid.arrange(p1, p2, p3, nrow=1,
   top = "Paired differences in Glucose level (before minus after diet)")
```

[1] http://teamcolorcodes.com/ is my source for team colors



We have two candidate outliers - one at -30 and one at 50. Based on a sample of 10 observations, it'd be hard to assume that these data were approximately Normally distributed. This suggests the use of a bootstrap confidence interval or a Wilcoxon signed rank test to make inferences about the population mean (or median) difference in glucose level.

Question 7i does not apply. These are paired samples.

Question 7j. Did a statistically significant ($\alpha = .05$) change occur after treatment? Include in your response evidence supporting any assumptions you make.

Based on a 95% bootstrap confidence interval, or a Wilcoxon signed rank test, we observe **no** statistically significant (at the 5% significance level) difference in population glucose levels before treatment as compared to after treatment.

Using the seed `43132`, I get the following 95% bootstrap confidence interval for the population mean of the paired before - after differences in glucose level.

```
set.seed(43132); Hmisc::smean.cl.boot(hw4q7$diffs)
```

```
   Mean    Lower    Upper 
 9.1000 -3.2025 21.2025 
```

From the Wilcoxon signed rank test, we have a p value that is substantially larger than the $\alpha = 0.05$ significance level we have established.

```
wilcox.test(hw4q7$diffs, conf.int = TRUE, conf.level = 0.95,
            exact = FALSE)
```

```
    Wilcoxon signed rank test with continuity correction

data:  hw4q7$diffs
V = 42, p-value = 0.1536
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 -6.999952 25.999955
sample estimates:
(pseudo)median
            10
```

The assumptions with the bootstrap are that we have a random (or at least representative) sample from the population of glucose differences, with those values drawn independently from an identical distribution. We don't require Normality of the distribution of that population.

Question 8 (50 points)

In the experiment studied in this question, 40 animals (20 male, 20 female) were randomly assigned either to "control" or to "histamine shock." The tabulated variable is medullary blood vessel surface (in mm^2/mm^3). Ignore the sex information.

```r
## Option 1 - create a hw4q8 data set in Excel, and import it into R
## using read.csv and tbl_df to make it into a tibble.
##
## Option 2 - build the tibble directly in R
hw4q8 <-
  tibble(
    animal = 1:40,
    sex = c(rep("F", 10), rep("M", 10), rep("F", 10), rep("M", 10)),
    trt = c(rep("Control", 20), rep("Histamine Shock", 20)),
    surface = c(6.4, 6.2, 6.9, 6.9, 5.4, 7.5, 6.1, 7.3, 5.9, 6.8,
                4.3, 7.5, 5.2, 4.9, 5.7, 4.3, 6.4, 6.2, 5.0, 5.0,
                8.4, 10.2, 6.2, 5.4, 5.5, 7.3, 5.2, 5.1, 5.7, 9.8,
                7.5, 6.7, 5.7, 4.9, 6.8, 6.6, 6.9, 11.8, 6.7, 9.0))
hw4q8
```

```
# A tibble: 40 x 4
   animal   sex     trt surface
    <int> <chr>   <chr>   <dbl>
 1      1     F Control     6.4
 2      2     F Control     6.2
 3      3     F Control     6.9
 4      4     F Control     6.9
 5      5     F Control     5.4
 6      6     F Control     7.5
 7      7     F Control     6.1
 8      8     F Control     7.3
 9      9     F Control     5.9
10     10     F Control     6.8
# ... with 30 more rows
```

Question 8a. Outcome (5 points)

The outcome under study is the medullary blood vessel surface measured in square mm per cubic mm.

Question 8b. Exposure/Treatment Groups (5 points)

The treatment was assigned as either "control" or "histamine shock."

Question 8c. Paired or Independent Samples (5 points)

These are independent samples. Each guinea pig was assigned to exactly one of the two treatments.

Question 8d. Random Sample? Representative? (5 points)

This isn't a random sample from the population of interest - which could be defined as all guinea pigs. The treatments were, however, assigned at random. Each animal was randomly assigned to their treatment, or, more likely, the assignment was done in such a way that all possible assignments of 20 animals to control and 20 to histamine shock were equally likely.

Question 8e. Significance Level? (5 points)

We are using a 95% confidence interval, so a 5% significance level.

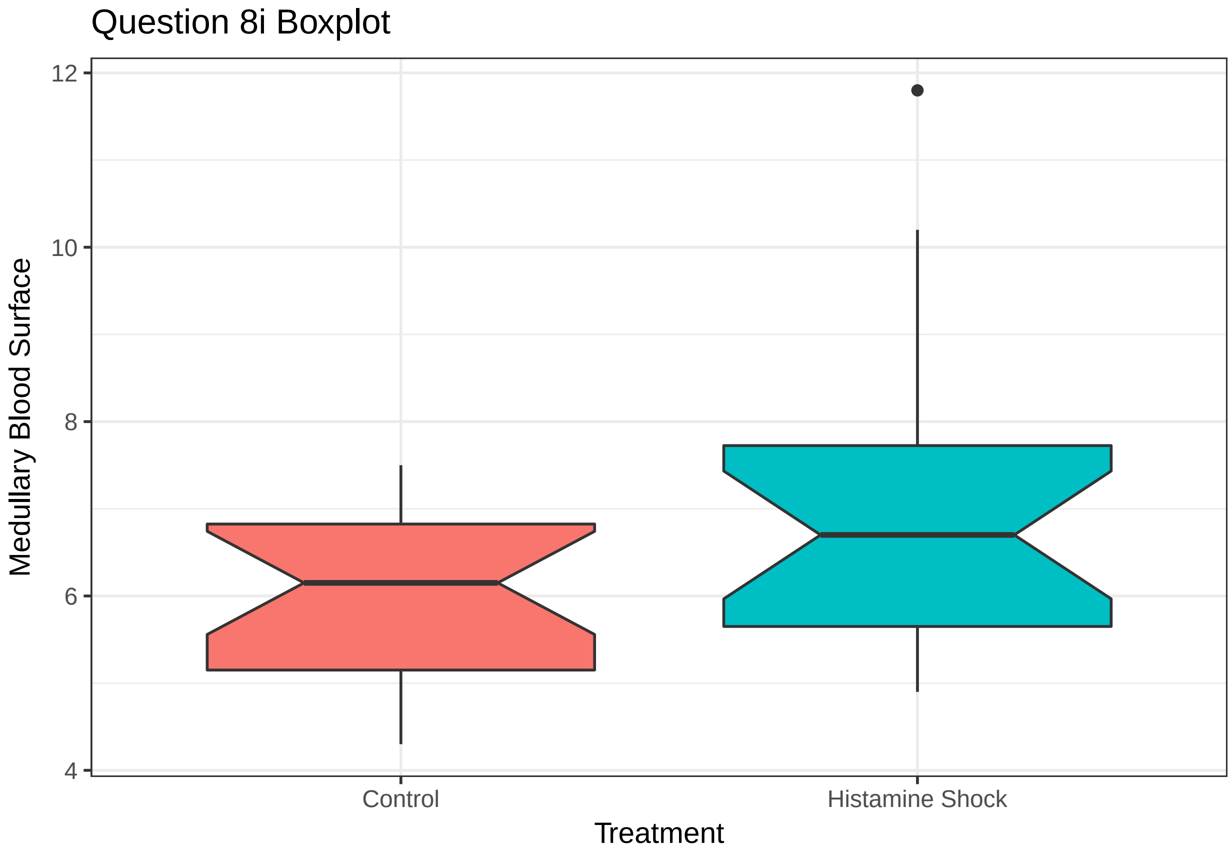
Question 8f. One-sided or Two-Sided Inference? (5 points)

We are looking for a significant difference between the two treatments, in either direction, so that's a two-sided approach.

Question 8i. What does the distribution of each individual sample tell us about which inferential procedure to use? (10 points)

Here is a comparison boxplot and a pair of normal Q-Q plots to compare the "control" and "histamine shock" surface results.

```r
ggplot(hw4q8, aes(x = trt, y = surface, fill = trt)) +
  geom_boxplot(notch = TRUE) +
  guides(fill = FALSE) +
  theme_bw() +
  labs(title = "Question 8i Boxplot", x = "Treatment",
       y = "Medullary Blood Surface")
```



We have a single outlier in the histamine shock group, at 11.8. Based on a sample of 20 observations, we'd expect to see an outlier, certainly, so that by itself isn't enough to push me away from assuming a Normal distribution for each of the two populations (control and histamine shock) here. I'd be inclined to use a confidence interval based on a t test. Given the balanced design with 20 animals in each group, whether or not we assume equal population variance shouldn't matter much.

Question 8j. On the basis of the most appropriate 95% CI, can you conclude that the outcome differs significantly between "control" and "histamine shock?" Justify your assumptions. (5 points)

Based on a 95% confidence interval using the pooled t procedure, we see a statistically significant (at the 5% significance level) difference between the control results and the results after histamine shock.

```
t.test(hw4q8$surface ~ hw4q8$trt, var.equal = TRUE)
```

```

    Two Sample t-test

data:  hw4q8$surface by hw4q8$trt
t = -2.2499, df = 38, p-value = 0.03033
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.0422568 -0.1077432
sample estimates:
        mean in group Control mean in group Histamine Shock
                        5.995                         7.070
```

The confidence interval (-2.04, -0.11) describes the control - histamine shock difference in surface, and does not include zero, although it almost does. The Welch t test gives a very similar confidence interval without requiring the assumption of equal variances, at (-2.05, -0.1). It turns out that the assumption of equal population variances is not very important here, in part because of the balanced design. That's the best answer here, in my view.

A slightly less good answer - the bootstrap

If, on the other hand, we had been unwilling to assume Normality in the two populations here, we would have to consider a bootstrap 95% confidence interval for the histamine shock - control differences. If we use a seed of `43163`, we get the following:

```
set.seed(43163); bootdif(hw4q8$surface, hw4q8$trt)
```

```
Mean Difference           0.025           0.975
       1.075000        0.239875        2.025000
```

Since this confidence interval does not contain zero, it also suggests that there is a statistically significant difference in mean medullary blood vessel surface between the two treatments.

A substantially less good answer - Wilcoxon-Mann-Whitney-based CI

Since we want to use a confidence interval here, I'd be inclined to stay away from the Wilcoxon-Mann-Whitney rank sum test, because it doesn't make a confidence interval about a difference in means, but rather a rather nebulous location variable. I use Wilcoxon-Mann-Whitney tests for *p* values, occasionally, but rarely do so when I want a confidence interval. If we decide to do it, we should probably turn off the `exact` part of the test, since we have some ties here.

```
wilcox.test(hw4q8$surface ~ hw4q8$trt, conf.int = TRUE,
            conf.level = 0.95, exact = FALSE)
```

```

    Wilcoxon rank sum test with continuity correction

data:  hw4q8$surface by hw4q8$trt
W = 138.5, p-value = 0.09854
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -1.7000064  0.1000035
sample estimates:
difference in location
             -0.699991
```

In this case, it matters if you do this test, because the confidence interval for the location parameter using that method, which is (-1.7, 0.1), does (barely) include 0. So, if we had used the Wilcoxon-Mann-Whitney procedure, we would have to conclude, barely, that the medullary blood vessel surface does not differ significantly between "control" and "histamine shock".

Setting up Questions 9-11

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
pander(zocazo[1:5,]) # see first five observations
```

subject	production	group
1	3.56	smoker
2	0.56	non-smoker
3	10.08	smoker
4	0.36	non-smoker
5	16.5	smoker

Question 9 (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(zocazo$production ~ zocazo$group, conf.lev=0.99, var.equal=TRUE)
```

```

    Two Sample t-test

data:  zocazo$production by zocazo$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.

Question 10 (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.

```
power.t.test(delta=0.2*(mean(zocazo$production[zocazo$group=="smoker"]) -
                          mean(zocazo$production[zocazo$group=="non-smoker"])),
             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.

Question 11 (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*(mean(zocazo$production[zocazo$group=="smoker"]) -
                          mean(zocazo$production[zocazo$group=="non-smoker"])),
             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*(mean(zocazo$production[zocazo$group=="smoker"]) -
                          mean(zocazo$production[zocazo$group=="non-smoker"])),
             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*(mean(zocazo$production[zocazo$group=="smoker"]) -
                          mean(zocazo$production[zocazo$group=="non-smoker"])),
             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
```