Emily Johnson

Senior Project

Dr. Shannon Pileggi

8 June 2018

How Correlated Data Affects Sample Size and Power Calculations

I Introduction:

The objective of this senior project was to learn about power and sample size calculations with correlated data. The motivation for this project was inspired by a researcher in the kinesiology department at Cal Poly, San Luis Obispo. The researcher, Dr. Ventura, wanted to examine the effectiveness of bottle feeding an infant using a clear bottle, which represents the control group, versus using an opaque bottle, which is the treatment group. The infant’s intake, measured in mL, was used as a way to assess if weight gain is lower or higher in the treatment group. Each infant’s intake was assessed at the baseline and at the four weeks follow up. Some of the infants were bottle fed with the clear bottle on the first visit and the opaque bottle on the second visit, while other infants were bottle fed with the opaque bottle on the first visit and the clear bottle on the second visit. The idea behind this was that if amount of milk left in the bottle was not visible, then possibly the infant would not finish the bottle and drink as much milk. The primary goal of this project was to use simulation to examine how the correlation in infant intake between the baseline and four weeks follow up affects the power and sample size calculations.

II. Data Exploration – Paired data

To begin my analysis, I started by diving the data into two separate groups. Group A is paired data that contains information for infants who were bottle fed with the clear bottle on the first visit and the opaque bottle on the second visit, and Group B is paired data that contains information for infants who were bottle fed with the opaque bottle on the first visit and the clear bottle on the second visit. After this, I computed some simple statistics on the data which are presented in the table below:

|  |  |  |
| --- | --- | --- |
|  | Group A | Group B |
| 1st visit mean intake | 98.37 mL (Clear) | 102.82 mL (Opaque) |
| 2nd visit mean intake | 89.49 mL (Opaque) | 109.79 mL (Clear) |
| 1st visit standard deviation | 53.70 mL | 44.47 mL |
| 2nd visit standard deviation | 44.66 mL | 49.45 mL |
| Mean difference | 8.88 mL | 6.97 mL |
| Standard deviation difference | 47.66 mL | 45.85 mL |
| Correlation between 1st and 2nd visit | 0.54 | 0.53 |
| Sample Size | 37 | 38 |

Table 1: Summary statistics of infant intake for Group A and Group B.

For both Group A and Group B, the correlation between infant intake on the first visit and the second visit is approximately 0.54. This means that there is a moderate, positive association between infant intake on the first visit and infant intake on the second visit. Also, the standard deviation for Group A and Group B is very large, 47. 66 mL and 45.75 mL respectively. Because of this, the power of the two, separate paired t-tests (one using Group A and one using Group B) is very low. Specifically, the power for Group A is only 19.6% and the power for Group B is only 14.8%. In fact, through simulation, I discovered that assuming all over values remain the same, we would need a sample size of at least 229 for Group A and a sample size of at least 342 to obtain 80% power. The next step in my investigation is determine if and how the correlation between the infant intake on the first and second visits plays a role in the power.

III. Power for a paired t-test

To further assess how correlation affects sample size calculations and the power of a test, I first created a new power function and then conducted a simulation. To do so, I manipulated the equation for the difference of the standard deviations of two groups in order to incorporate correlation, which I have shown below:

cor(X,Y) = cov(X,Y)/(sd(X)\*sd(Y))

cov(X,Y) = cor(X,Y)\*sd(X)\*sd(Y)

var(X−Y) = var(X)+var(Y)−2\*cov(X,Y)

var(X−Y) = var(X)+var(Y)−2\*cor(X,Y)\*sd(X)\*sd(Y)

sd(X−Y) = sqrt(var(X) + var(Y) – 2\*cor(X,Y)\*sd(X)\*sd(Y))

Because standard deviation influences power calculations, this manipulation of the equation allows us to assess how correlation will influence power. The power function that already exists within R takes the parameters of sample size, the mean of the differences between the two groups, the standard deviation of the differences, and the type of test, which is paired for this data. The new function that I created using R takes the parameters of the two standard deviations of the separate groups as well as the correlation, in addition to sample size and mean of the differences. This allows us to use the equation that I manipulated above to solve for the standard deviation of the differences while incorporating the correlation. The code for the function to calculate power for paired data is below:

power\_func <- function(r, n, s1, s2, delta)

{

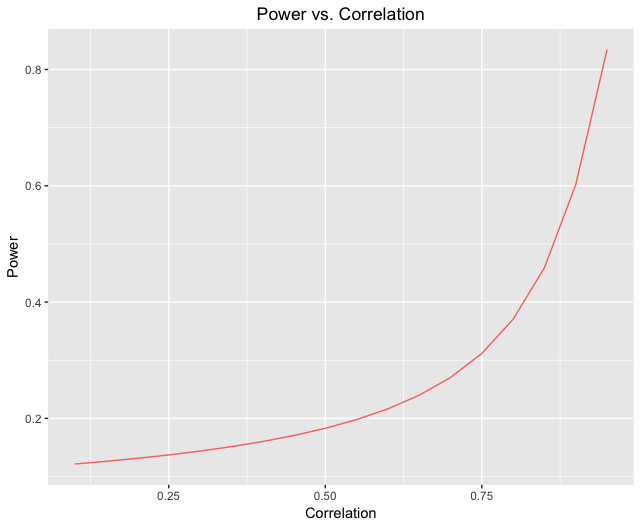
sd\_diff <- sqrt(s1^2 + s2^2 - 2\*r\*s1\*s2)

power <- power.t.test(n, delta = delta, sd = sd\_diff, type = "paired")

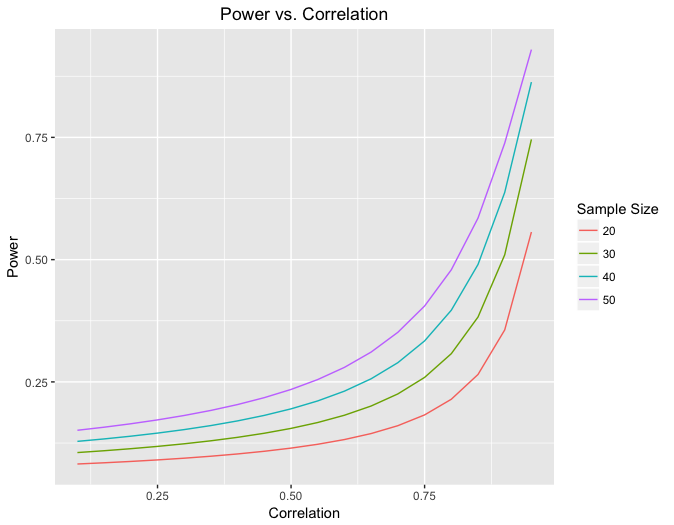
return(power$power)

}

Now, we can assess how much the correlation truly impacts the power and sample size calculations. I investigated this question separately for Group A and Group B. For Group A, the power was initially 19.58%. If we increase our correlation to .80, keeping all other values the same, the power roughly doubles to 37.05%. Similarly, for Group B, the power was initially 14.75%. If we increase the correlation to .80, the power also roughly doubles to 28.53%. If we go back and look at the equation we manipulated to solve for the standard deviation of the differences between visit 1 and visit 2, when we increase the correlation, that results in the standard deviation of the differences decreasing. Thus, as the standard deviation of any test decreases, the power will increase. Therefore, we can assume that as correlation increases, power also increases. Graph 1 depicts a visual representation of the relationship between power and correlation, specifically when using values that coordinate to the Group A data. Graph 2 below depicts the same relationship between power and correlation, except for altering the sample size while still keeping all other values the same:

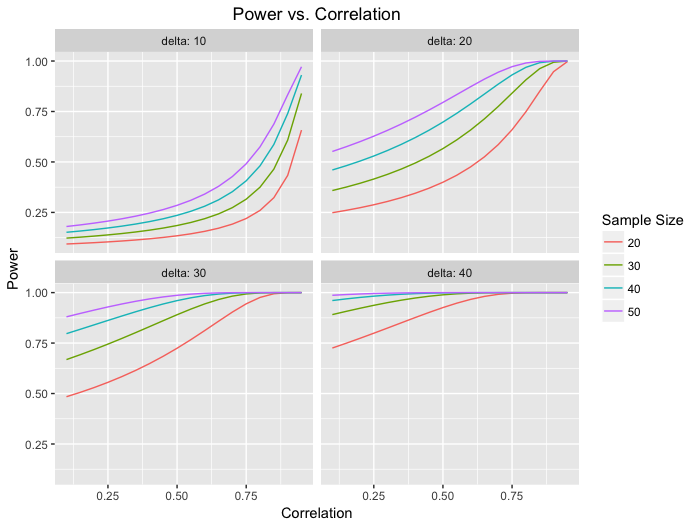


Graph 1



Graph 2

The positive relationship between sample size and power in visible here; as we increase sample size, power increases as well. Lastly, lets alter another variable. Graph 3 depicts the same relationship as Graph 2, however, separated by varying values for the difference of means, which is labeled as delta:



Graph 3

Here you can see the relationship that both correlation and delta play into calculating power. As our difference in means between visit 1 and visit 2 is increasing, the power increases. If we have a very high difference in means in relation to the standard deviation and a very high correlation, as depicted in Graph 3, the power increases closer to 1.

IV. Power for a two-sample t-test

The next step of my investigation was inspired by the hypothetical idea of testing if one group of infants was given both the clear bottle on their first and second visits, and the second group of infants was given both the opaque bottle on their first and second visits, what would be the difference in infant intake between those two groups? Now, we essentially have a two-sample test with two groups of paired data. Hence, this calculation is a bit more complicated. The first step was to create a similar power function to the one I had written previously, but taking into account the additional parameters needed. Also, there is not an accurate power function already embedded within R to properly calculate the power of a two-sample t-test, therefore that component had to be written into the function as well. The function is coded below:

power\_2samptest <- function(n1, n2, s1a, s1b, s2a, s2b, r1, r2,

delta1, delta2, pooled)

{

sd1\_diff <- sqrt(s1a^2 + s1b^2 - 2\*r1\*s1a\*s1b)

sd2\_diff <- sqrt(s2a^2 + s2b^2 - 2\*r2\*s2a\*s2b)

if(pooled == TRUE)

{

std\_error <- (sqrt(((n1 - 1)\*sd1\_diff^2 + (n2 -

1)\*sd2\_diff^2)/(n1 + n2 - 2)))\*sqrt(1/n1 + 1/n2)

}

else

{

std\_error <- sqrt(sd1\_diff^2/n1 + sd2\_diff^2/n2)

}

df <- n1 + n2 - 2

mu0 <- 0

reps <- 10000

outsideCI <- numeric(reps)

set.seed(2)

for (i in 1:reps) {

x <- rnorm(n1, delta1, sd1\_diff)

y <- rnorm(n2, delta2, sd2\_diff)

CI.lower <- (mean(x)-mean(y)) - qt(0.975, df)\*std\_error

CI.upper <- (mean(x)-mean(y)) + qt(0.975, df)\*std\_error

outsideCI[i] <- ifelse(mu0 < CI.lower | mu0 > CI.upper, 1,

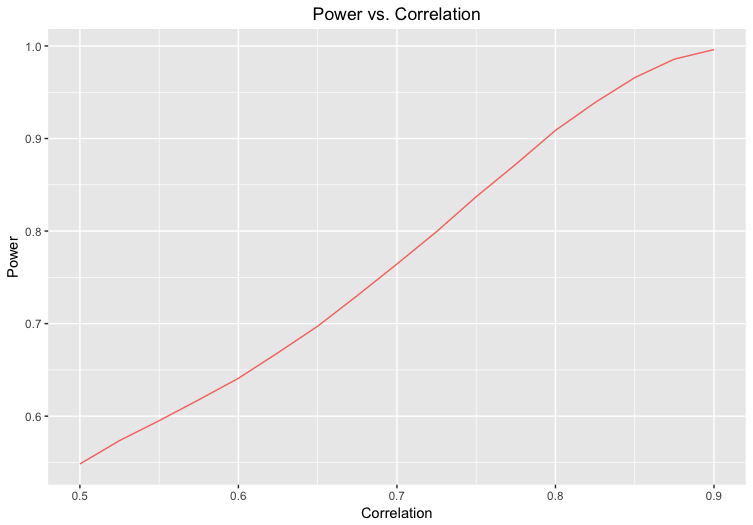
0)

}

return(mean(outsideCI))

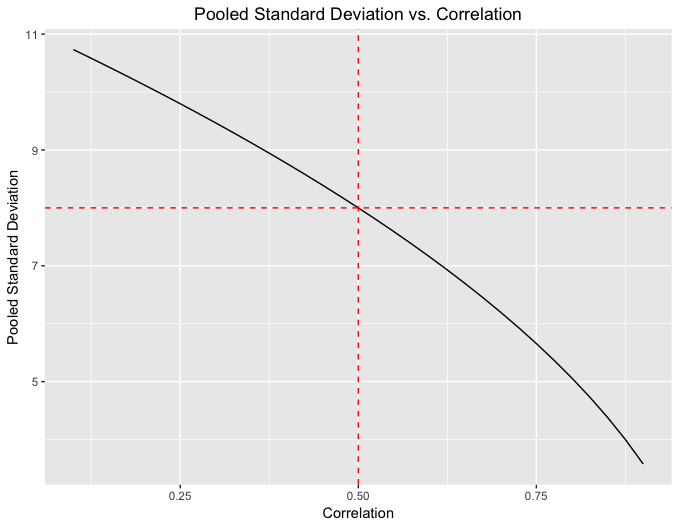
}

This function essentially takes the same parameters as the previous function, times two. There are now two sample sizes, two correlations, two deltas, and two standard deviations for each group, totaling four different standard deviation values. Also, the function has a TRUE/FALSE argument named “pooled”. If we have good reason to believe that the variances from the two populations are about the same, then the argument would be set to “TRUE”. If we do not believe that to be the case, then the argument would be set to “FALSE”. Similar to the simpler paired sample calculations, as correlation increases, power does as well, which can be seen in Graph 4 below:



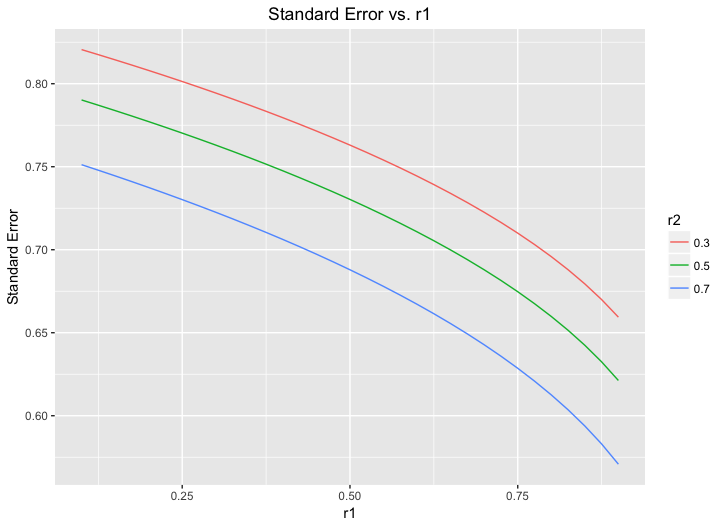
Graph 4

Graph 4 is assuming that both samples have a sample size of 35, the standard deviations of visit 1 and visit 2 for both samples is 8 (pooled = TRUE), and the difference in means of the two samples is 4. Now, let’s assume that the standard deviations of visit 1 and visit 2 are the same for both samples, therefore pooled is TRUE. If the correlation for both samples is less than 0.5, then the standard deviation of the differences for both samples will be greater than the standard deviation of visit 1 and visit 2. As we increase the correlation closer to 0.5, the standard deviation of the differences for sample 1 and sample 2 and the pooled standard deviation will get smaller and reach closer to the standard deviation of visit 1 and visit 2. As a result, as the correlation increases, the pooled standard deviation gets smaller, and thus power increases. On the other hand, as we decrease the correlation further away from 0.5, the standard deviation of the differences for sample 1 and sample 2 and the pooled standard deviation will still get smaller, however it is moving further away from the standard deviation of visit 1 and visit 2. Graph 5 demonstrates this idea. The vertical dotted line is plotted when the correlation is equal to 0.5, and therefore the pooled standard deviation is the same as the standard deviation from visit 1 and visit 2, which is 8 mL. As the correlation decreases from 0.5, the pooled standard deviation gets larger and as the correlation increases, the pooled standard deviation gets smaller. As a result, as the pooled standard deviation gets smaller, the power increases.

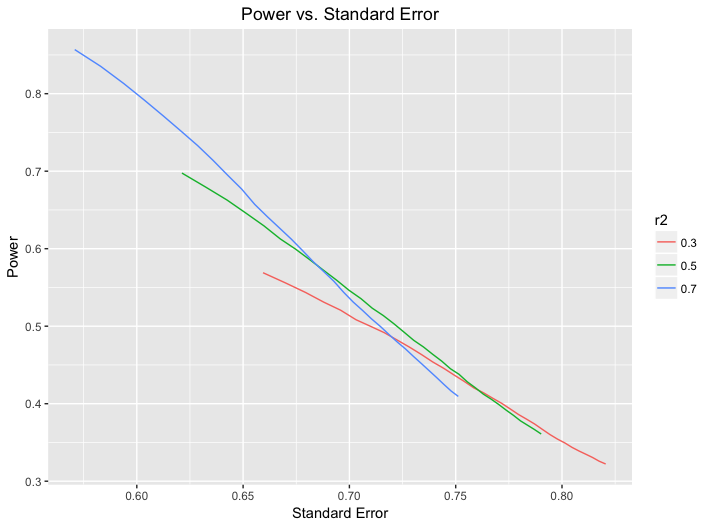


Graph 5

Graph 6 is assuming that the sample size for both samples is 30, the difference in means of the two samples is 4, and the standard deviation of visit 1 and visit 2 for both samples is 8. Here, the correlation of group 1 increases from 0.1 to 0.9 by 0.025 and the correlation of group 2 is either 0.3, 0.5, or 0.7. This graph displays that for the stronger values of correlation for both samples, the standard error is the smallest. In contract, for the weakest values of correlation for both samples, the standard error is the largest. In this case, pooled is set to FALSE because the correlations are different for both samples, therefore we are using a standard error instead of a pooled standard deviation. Using the same data as in Graph 6, Graph 7 displays how power is affected by the standard error. As the standard error is increasing, the power is decreasing. This makes intuitive sense because when computing the t statistic for a two-sample t test, the standard error is on the bottom of the fraction. Therefore, for larger values of standard error, the t statistic is smaller, and the smaller the t statistic, the smaller the power.



Graph 6



Graph 7

I would like to thank Dr. Ventura for providing us with the data that inspired this exploration of how correlated data affects sample size and power calculations. The data given to us was in the form of paired data, which provided us with a basis for the first part of my analysis. To go a bit further, we also looked at the case of two-sample data, both samples containing paired data. The results of the simulation and functions I created in R revealed, at the simplest level, that the more strongly correlated your data is, the higher your power will be. The reason for this is because correlation plays a role in the standard deviation of the differences between two groups, in our case, visit 1 data and visit 2 data. As we increase correlation, the standard deviation of the differences decreases, and as we already know, a smaller standard deviation results in higher power. Correlation is an important aspect to keep in mind when designing a study. If you know that you are going to be collecting data that is more highly correlated, you can be less conservative with your design to still achieve a certain set value of power.