POL SCI 231b (Spring 2017): Problem Set 1

Prof. Thad Dunning/GSI Natalia Garbiras-Díaz

Dept. of Political Science

University of California, Berkeley

Due Friday, January 27 in section

Please email only **one** copy of your solution set per group, with the names of all group members on it. Please use the following email address: nataliagarbirasdiaz+231b@berkeley.edu. If you need to turn in a hard copy, please do so in section and only turn in one per group.

Remember to work out the problems on your own, before you meet with your group to agree on solutions.

1. **(From Freedman, Pisani, and Purves—hereafter FPP). A gambler will play roulette 50 times, betting a dollar on four joining numbers each time (like 23, 24, 26, 27 in FPP 4th edition, figure 3, p. 282). If one of these four numbers comes up, she gets the dollar back, together with winnings of $8. If any other number comes up, she loses the dollar. So this bet pays 8 to 1, and there are 4 chances in 38 of winning. Her net gain in 50 plays is like the sum of ___ draws from the box ___. Fill in the blanks; explain.**

   Her net gain is like the sum of 50 draws from a box with 4 tickets marked “$8” and 34 tickets marked “-$1.” Explanation: we construct the box so that draws reflect the relative chances of the different outcomes occurring—and the outcomes reflect the net gain in each case.

2. **(From FPP). You are thinking about playing a lottery. The rules: you buy a ticket, choose 3 different numbers from 1 to 100, and write them on the ticket. The lottery has a box with 100 balls numbered from 1 through 100. Three balls are drawn at random without replacement. If the numbers on the balls are the same as the numbers on your ticket, you win. (Order doesn't matter). If you decide to play, what is your chance of winning? Briefly explain your answer.**

The probability that the first number drawn is any of three numbers x, y, or z (e.g., 1, 2, or 3) is $3/100$. If one of these numbers is drawn, the probability that the second draw turns up one of the remaining two numbers is $2/99$. Finally, if two of the numbers are drawn on the first two draws, the probability that the last draw is the remaining number is $1/98$. Thus, the probability that all three numbers are drawn is $(3/100) * (2/99) * (1/98) = 6/970,200$.

Notes: in the background, you are using the rule of summing the probabilities of mutually exclusive events: for example, the probability that the first ball is x, y, OR z (each mutually exclusive events) is the sum of their individual probabilities: $1/100+1/100+1/100 = 3/100$. After you have the probabilities for each draw, you multiply the conditional probabilities. If this is unfamiliar, read FPP, Chapters 13-14.

3. **(From FPP).**

   (a) **Four draws are going to be made at random with replacement from a box with five tickets in it. The tickets are labelled 1, 2, 2, 3, 3, respectively. Find the chance that 2 is drawn at least once.**

   On each draw, 2 is drawn with probability $2/5$; with probability $3/5$, a different number is drawn. Thus, the probability that a different number appears on each of four draws is $(3/5)^4 = 81/625 \doteq 0.13$. (Remember, the draws are independent as we are drawing with replacement, so you can multiply the probabilities). Thus, the chance that 2 appears on at least one of the four draws is $1 - 0.13 = 0.87$.

   Note: to figure out the chance of an event, it is sometimes useful to figure out the chance of its opposite; then subtract from 100%.

   (b) **Repeat (a), if the draws are made at random without replacement.**

   2 must be drawn after four draws: the probability is 1.

4. **There are currently 12 students registered for our class. (a) Suppose I take a survey of the 12 students to find out who is left-handed. Is the number of left-handers a random variable?**

   A random variable is a chance procedure for generating a number—such as drawing tickets at random from a box. When we draw tickets, we can describe the chances that different numbers will be drawn, based on what is in the box. (In practice, we may not know what's in the box—but the unobserved values of the tickets in the box still determine the chances).

   Here, there's no such chance procedure. We are interviewing all 12 students in the class: it is as if we are observing every ticket in the box. There is no chance procedure that assigns students to be left-handed. Nor are we drawing the 12 class members at random from a larger well-defined population. Thus, we can't readily describe the chances of different outcomes occurring in terms of the features of some underlying box.

   (b) Suppose I find that there are 8 right-handers and 4 left-handers. Is the difference statistically significant? Explain your answers. (Note: we are covering hypothesis

testing on Thursday 1/26, but you may already have the knowledge you need to answer this question.).

The goal of significance testing is to assess whether certain observed statistics could reasonably arise by chance, given some hypothesis about what's in the box. For example, one might ask: if we hypothetically repeated a given chance procedure over and over again, how often would we observe an observed statistic as extreme as the one we in fact observed—if our hypothesis about what's in the box is correct?

However, significance testing doesn't make sense without random variables. In our example, there's no well-defined chance procedure that can hypothetically get repeated over and over again. For instance, we are cannot readily ”assign” students over and over again to be right-handed or left-handed. Nor are we sampling our class from a population, in a way that can be described using a box model. The broader lesson: without a realistic box model, significance testing doesn't make sense.

5. **Consider a randomized experiment in which N units are assigned to a treatment or control group. Denote the m units assigned to treatment by $i = 1, ..., m$, with $Y^T = \frac{\sum_i^m Y_i}{m}$, and the $N - m$ units assigned to control by $i = m + 1, ..., N$, with $Y^C = \frac{\sum_{m+1}^N Y_i}{N-m}$. Here, Y_i is an observed outcome. Let $\overline{Y(1)} \equiv \sum_i^N \frac{Y_i(1)}{N}$ and $\overline{Y(0)} \equiv \sum_i^N \frac{Y_i(0)}{N}$. (Note that $\equiv$ means "identically equal to.") Now, answer the following questions and briefly explain your answers:**

   (a) **What is the quantity $\overline{Y(1)} - \overline{Y(0)}$? Is this an estimator and if so, what does it estimate?**

   This is the average causal effect. It is a parameter (not an estimator) and refers to the difference between the means of the potential outcomes under treatment and control.

   (b) **What is the quantity $\frac{1}{N} \sum_i^N (Y_i(1) - \overline{Y(1)})^2$? Is this an estimator and if so, what does it estimate?**

   This is the variance of the potential outcomes under treatment. It is a parameter.

   (c) **What is the quantity $Y^T - Y^C$? Is this an estimator and if so, what does it estimate?**

   $Y^T - Y^C$ is the difference of the means of the units assigned to treatment and control. It is an estimator of $\overline{Y(1)} - \overline{Y(0)}$, the average causal effect.

   (d) **For each of the estimators you identified in parts (a)-(c), show that the estimator is unbiased for its estimand (i.e., the parameter it estimates).**

   We only identified one estimator, the difference in means in (c). Estimators use calculations based on sample data—statistics—to estimate unknown features of the population from which a sample is drawn—parameters. Here, the parameter is the average causal effect: the difference between the average potential outcome under treatment and the average potential outcome under control, where the average is taken over the whole study group. (That's why this is a parameter: it characterizes the population from which the sample is drawn). In notation, this parameter is

$$\tau = \frac{1}{N}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)]. \tag{1}$$

An estimator is unbiased if its expectation equals the parameter. Here, our estimator is the difference of means, $Y^T - Y^C$, that is, the difference between the average outcome in the treatment group (the treatment sample) and the average outcome in the control group (the control sample). And the parameter is τ.

Thus, the difference of means is an unbiased estimator for τ if $E[Y^T - Y^C] = \tau$. To see that the difference-of-means estimator is unbiased, note first that by substitution,

$$E[Y^T - Y^C] = E[\frac{\sum_{i=1}^{m} Y_i}{m} - \frac{\sum_{i=m+1}^{N} Y_i}{N-m}].$$

Then, distributing expectations, we have

$$\begin{aligned}
E[Y^T - Y^C] &= E[\frac{\sum_{i=1}^{m} Y_i}{m}] - E[\frac{\sum_{i=m+1}^{N} Y_i}{N-m}] \\
&= \frac{\sum_{i=1}^{m} E(Y_i)}{m} - \frac{\sum_{i=m+1}^{N} E(Y_i)}{N-m} \qquad (2)\\
&= \frac{m[\frac{1}{N}\sum_{i=1}^{N} Y_i(1)]}{m} - \frac{(N-m)[\frac{1}{N}\sum_{i=1}^{N} Y_i(0)]}{N-m} \qquad (3)\\
&= \frac{1}{N}\sum_{i=1}^{N}[Y_i(1)] - \frac{1}{N}\sum_{i=1}^{N}[Y_i(0)] \\
&= \tau.
\end{aligned}$$

There are two key ideas in this derivation. First, in (2), we can distribute expectations since we are drawing at random: the unconditional expectation of each draw is the same, whether $i = 1$ or $i = 2$ or $i = ...$

Second, in (3), the expected value of each draw Y_i is the average of the potential outcomes (the mean of the box). So we can plug in $\frac{1}{N}\sum_{i=1}^{N} Y_i(1)$ for $E(Y_i)$ in the treatment group and $\frac{1}{N}\sum_{i=1}^{N} Y_i(0)$ for $E(Y_i)$ in the control group. Since these are constants, the summation just gives $m * [\frac{1}{N}\sum_{i=1}^{N} Y_i(1)]$ for the first term and $(N - m) * \frac{1}{N}\sum_{i=1}^{N} Y_i(0)$ for the second term, in the numerator of (3).

6. **Suppose we draw $i = 1, ..., n$ tickets at random with replacement from a box. The mean of the tickets in the box is μ, and their variance is σ^2. Denote the value of a given draw of a ticket by Y_i and the average of all n tickets by $\overline{Y}$. For each of the following statements, say which aspects of this sampling process must be invoked for the statement to true. Explain your answers, using English as well as algebra (where appropriate).**

(a) $E(Y_1) = \mu$

Random sampling.

(b) $E(Y_2) = \mu$

Random sampling.

(c) $E(\overline{Y}) = \mu$

With random sampling, the expected value of the sample mean equals the population mean. You could also derive this directly:

$$\begin{aligned}
E(\overline{Y}) &= E(\frac{\sum_{i=1}^{n} Y_i}{n}) \\
&= \frac{\sum_{i=1}^{n} E(Y_i)}{n} \qquad (4) \\
&= \frac{n * \mu}{n} \qquad (5) \\
&= \mu.
\end{aligned}$$

In (4), we distributed expectations; in (5), we used the result from parts (a) and (b).

(d) $\text{Var}(Y_i) = \sigma^2$

The variance of a single draw is the variance of the box. Y_i is a random variable so the variance of Y_i is equal to the standard error, squared. (We often talk about the "standard error" to mean "the standard error of the mean," but the term is more general). Implicitly, we are also invoking the fact that we are drawing with replacement (and thus the draws are identically distributed): the variance of the box is the same for any draw. The distribution of the box does not change as we go.

(e) $\text{Var}(\overline{Y}) = \frac{\sigma^2}{n}$

The sampling variance of the sample mean is the variance of the box, divided by the number of tickets in the sample. You can derive this directly:

$$
\begin{aligned}
\text{Var}(\overline{Y}) &= \text{Var}(\frac{\sum_{i=1}^{n} Y_i}{n}) \\
&= \frac{\sum_{i=1}^{n} \text{Var}(Y_i)}{n^2} \\
&= \frac{n * \sigma^2}{n^2} \\
&= \frac{\sigma^2}{n}
\end{aligned}
\qquad (6)
$$

Here, we need to invoke the property of drawing with replacement, which allows us to distribute the variance in (6): the draws are identically distributed, so the variance of each draw is the same.

(f) $Y_i \sim$ i.i.d. :

The Y_is are independent because we are drawing with replacement. For the same reason, they are identically distributed: we are drawing over and over again from the same population. The box doesn't change as we go, because we put the tickets back in each time.

(g) Y_i **is a random variable**

Y_i is a random variable because it the outcome of a chance process, i.e. of random sampling from a box.

7. **Now suppose we draw** n **tickets at random without replacement. Which of the statements in parts (a)-(g) of the previous question are true and which are false? Explain your answers.**

(a) True. The expected value of the first draw is the mean of the box.

(b) True. Unconditionally, the expected value of the second draw is the mean of the box.

(c) True. The expected value of the sample mean is the mean of the box.

(d) True: the variance of a single draw (a random variable) is the variance of the box. Even though drawing without replacement—so the distribution of the box changes as we go—the unconditional variance of any single draw is the same as before.

(e) Not true. Since we are sampling without replacement, the variance of $\overline{Y}$ is going to be smaller relative to sampling with replacement: the sample looks more and more like the population as we go. (Imagine that we drew all but one ticket from the population for each sample: the means would not vary much, across hypothetical replications of the sampling process; they would be very close to the mean of the box.) Thus, we need a finite-sample correction factor: if we are drawing n tickets without replacement from a box of N tickets, the sampling variance of the mean would be

$$\frac{N-n}{N-1}\frac{\sigma^2}{n}. \tag{7}$$

This is smaller than $\frac{\sigma^2}{n}$ because $\frac{N-n}{N-1} < 1$ as long as we are drawing more than one ticket. (Note: derivation of the finite-sample correction factor is a topic for another day).

(f) Not true: the draws are no longer independent, nor identically distributed.

(g) True: Y_i is still a random variable.

8. **(R exercise) Consider the dataset called "potential outcomes" that is loaded at the Problem Set 1 folder on the class bCourses page. Write and turn in R code that does the following (please comment your code to indicate what each line or chunk of code is doing), and answer the conceptual questions as comments in your code:**

   - **Find the difference in average potential outcomes under treatment and control, for the study group. What is another term for this quantity? Is this an estimator or a parameter, and why?**

     Average causal effect (ACE). This is a parameter.

     ```r
     mean(data$Y_i1) - mean(data$Y_i0)

     ## [1] 3.717696
     ```

   - **Simulate an experiment in which 1/2 of the units are assigned to the treatment group and 1/2 are assigned to control. Calculate the difference of means between the treatment and control group.**

     ```r
     # First we will create a fake treatment vector
     # assigning half of the units to treatment and
     # half to the control, we can then sample
     # without replacement from this vector to get
     # other fake treatments with the same number
     # of units assigned to treatment and control.

     treat <- c(rep(1, nrow(data)/2), rep(0, nrow(data)/2))
     ```

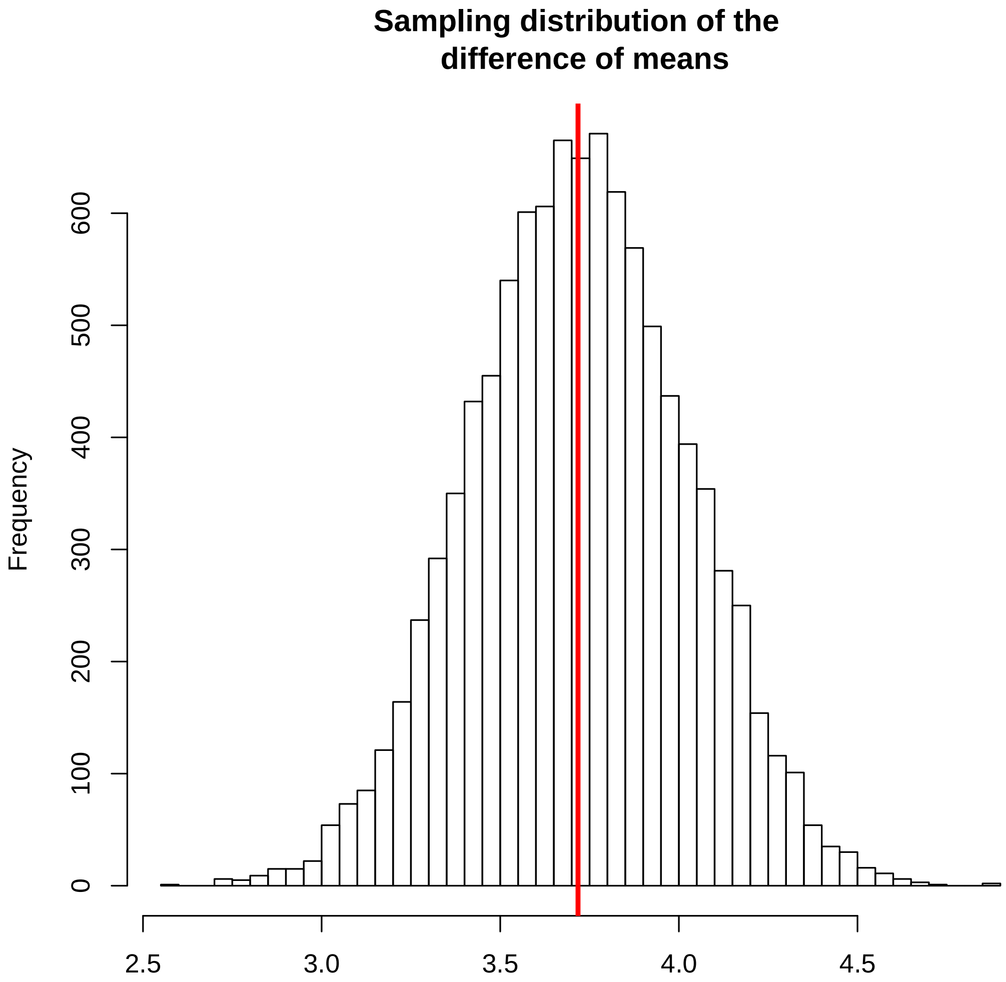
```
fake_treat <- sample(treat, length(treat), replace = FALSE)
table(fake_treat)

## fake_treat
##  0  1
## 23 23

# Now we calculate the difference of means
mean(data$Y_i1[fake_treat == 1]) - mean(data$Y_i0[fake_treat ==
    0])

## [1] 3.250043
```

- **Replicate step (2) 10,000 times, saving the difference of means for each replicate.**

```
# placeholder vector to put all the
# differences
dm <- NA
# and now we write the loop
for (i in 1:10000) {

    fake_treat <- sample(treat, length(treat),
        replace = FALSE)

    dm[i] <- mean(data$Y_i1[fake_treat == 1]) -
        mean(data$Y_i0[fake_treat == 0])

}
```

- **Plot a histogram for the 10,000 differences of means, and add a vertical line showing the average causal effect you calculated in (1).**

```
hist(dm, breaks = 50, main = "Sampling distribution of the \n difference of me
    xlab = "")
# and we add a red line for the ACE
abline(v = mean(data$Y_i1) - mean(data$Y_i0),
    col = "red", lwd = 3)
```



- **Is the difference of means in 2. a parameter or an estimator? If the latter, what does it estimate, and does the evidence suggest that it is biased or unbiased? Explain your answer.**

  The difference in means is an estimator of the average causal effect. The histogram suggests it is unbiased: the sampling distribution of the difference in means is centered around the ACE.

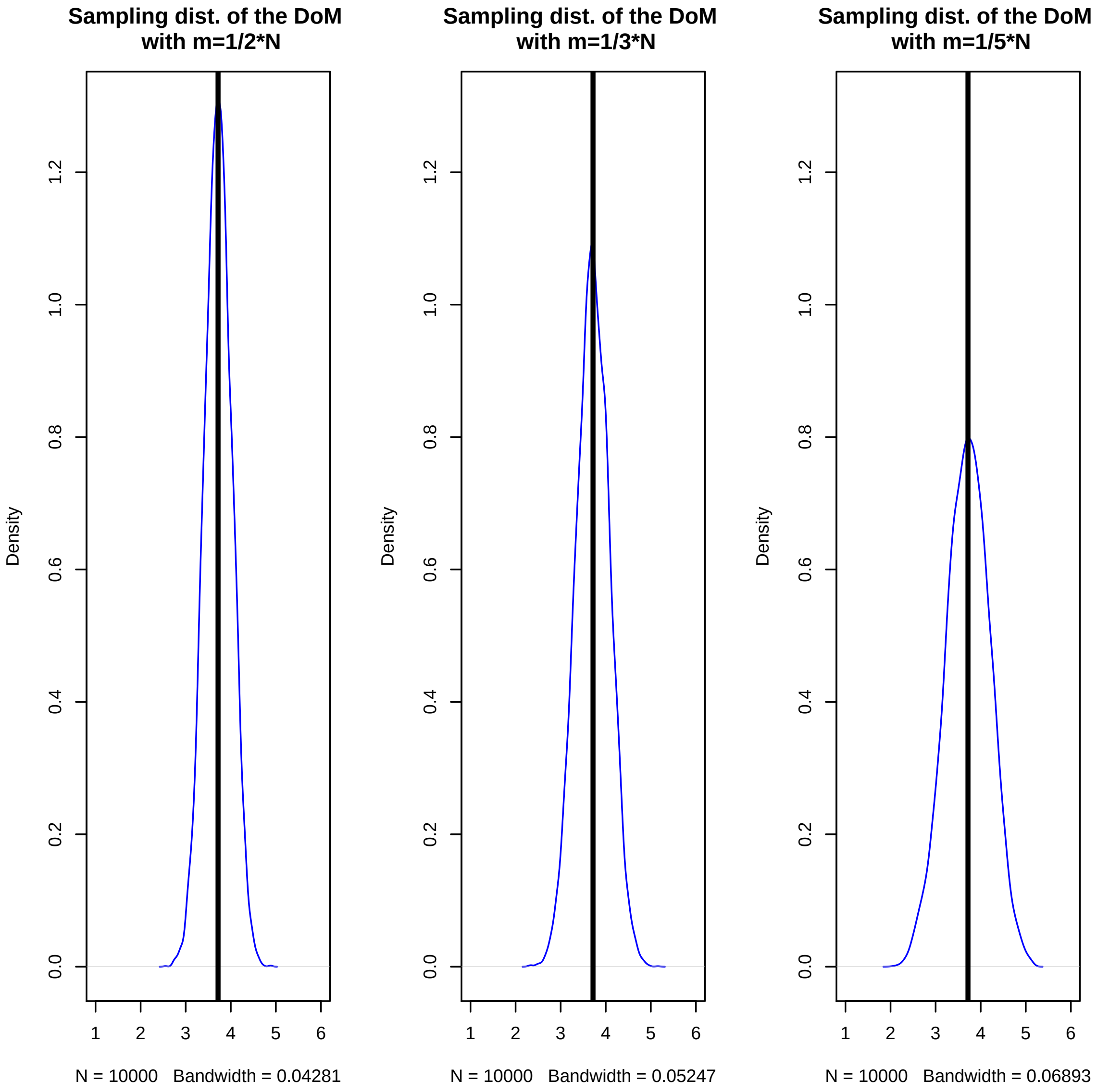
- **What is the standard error of the difference of means? Use your code and plot to answer this question, and explain your answer.**

  The standard error of the difference of means is the SD of the sampling distribution of the difference in means. Thus,

```
sqrt(mean((dm - mean(dm))^2))

## [1] 0.3001292
```

- **Repeat the simulation in the previous exercise but simulating experiments that vary the proportion of units in the treatment and control group.**
  - 1/5 of the units are assigned to treatment, 4/5 to control.
  - 1/3 of the units are assigned to treatment, 2/3 to control.

```
# To do this, we will generate new treatment
# vectors that assign this number of units to
# treatment/control.

round(nrow(data)/5)  #1/5 of units assigned to the treatment group

## [1] 9

treat.1.5 <- c(rep(1, round(nrow(data)/5)), rep(0,
    46 - round(nrow(data)/5)))
treat.1.5

##  [1] 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0
## [21] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [41] 0 0 0 0 0 0

round(nrow(data)/3)  #1/3 of units assigned to the treatment group

## [1] 15

treat.1.3 <- c(rep(1, round(nrow(data)/3)), rep(0,
    46 - round(nrow(data)/3)))
treat.1.3

##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0
## [21] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [41] 0 0 0 0 0 0

# and now we write the loop first for the case
# with 1/5 of the observations in the control
# group placeholder to put all the differences
dm.1.5 <- NA

for (i in 1:10000) {

    # fake treatment vector
    fake_treat <- sample(treat.1.5, length(treat),
        replace = FALSE)

    # and calculate the difference of means
    dm.1.5[i] <- mean(data$Y_i1[fake_treat ==
        1]) - mean(data$Y_i0[fake_treat == 0])

}

# and now with 1/3 of the observations in the
# control group placeholder to put all the
# differences
dm.1.3 <- NA

for (i in 1:10000) {

    # fake treatment vector
    fake_treat <- sample(treat.1.3, length(treat),
        replace = FALSE)

    # and calculate the difference of means
    dm.1.3[i] <- mean(data$Y_i1[fake_treat ==
        1]) - mean(data$Y_i0[fake_treat == 0])

}

par(mfrow = c(1, 3))  #including three plots in the same line
plot(density(dm), lwd = 1, col = "blue", ylim = c(0,
    1.3), xlim = c(1, 6), main = "Sampling dist. of the DoM \n with m=1/2*N")
abline(v = mean(data$Y_i1) - mean(data$Y_i0),
    lwd = 3)
plot(density(dm.1.3), lwd = 1, col = "blue", ylim = c(0,
    1.3), xlim = c(1, 6), main = "Sampling dist. of the DoM \n with m=1/3*N")
abline(v = mean(data$Y_i1) - mean(data$Y_i0),
    lwd = 3)
plot(density(dm.1.5), lwd = 1, col = "blue", ylim = c(0,
    1.3), xlim = c(1, 6), main = "Sampling dist. of the DoM \n with m=1/5*N")
abline(v = mean(data$Y_i1) - mean(data$Y_i0),
    lwd = 3)
```



The simulation shows that the difference in means is centered around the true value for the average causal effect for all cases. However, the spread of the distribution varies with the proportion of units assigned to treatment and control. The variance of the sampling distribution is minimized when the proportion of units in the treatment and control is the same. However, this will depend on the variance of the potential outcomes under treatment and control are about the same.