POLI SCI 231b: Problem Set 2 Solution Set

Spring 2017

University of California, Berkeley

Prof. Thad Dunning/GSI Natalia Garbiras-Díaz

1. **Let X be a variable with average $\overline{X}$. Using the definition of variance, show using math notation that for any constant m, $\text{Var}(\frac{X}{m}) = \frac{\text{Var(X)}}{m^2}$.**

$$\begin{aligned}
Var(\frac{X}{m}) &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i}{m} - \frac{1}{N}\sum_{i=1}^{N}\frac{X_i}{m}\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{m} * (X_i - \frac{1}{N}\sum_{i=1}^{N}X_i)\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{m^2}(X_i - \bar{X})^2\right) \\
&= \frac{1}{m^2}\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2 \\
&= \frac{1}{m^2}Var(X).
\end{aligned}$$

2. **Draws are being made at random with replacement from a box. The number of draws is getting larger and larger. Say whether each of the following statements is true or false, and explain. (Remember that "converges" means "gets closer and closer.")**

   (a) **The probability histogram for the sum of the draws (when put in standard units) converges to the standard normal curve.**
   True. The 'probability histogram' is the theoretical, or ideal histogram of chances (FPP, chapter 18). This is a synonym for "probability distribution" and "sampling distribution." When the number of draws is large, the probability histogram for the sum of the draws, when put in standard units, will converge to the standard normal curve (by the central limit theorem).

(b) **The histogram for the numbers in the box (when put in standard units) converges to the standard normal curve.**

False. If the numbers in the box are normally distributed, then the histogram for the numbers in the box follows the normal curve. (Note that standardizing a normal distribution by subtracting the mean and dividing by the standard deviation gives a standard normal distribution). If the numbers are not normally distributed, the histogram does not follow the normal curve. The numbers do not "converge to" the standard normal curve. The distribution of the box is what it is; it is unaffected by the size of the sample.

(c) **The histogram for the numbers drawn (when put in standard units) converges to the standard normal curve.**

False. The distribution of the numbers drawn will converge to the distribution of the box.

(d) **The probability histogram for the product of the draws (when put in standard units) converges to the standard normal curve.**

False. The normal curve is tied to sums, not products (FPP 18.5). That is, the central limit theorem does not apply to the product, only the sum and average.

(e) **The histogram for the numbers drawn converges to the histogram for the numbers in the box.**

True. The empirical histogram of the numbers drawn will look more and more like the histogram of the numbers in the box as n increases.

(f) **The variance of the numbers drawn converges to zero.**

False; see (e). If the numbers in the box have a positive variance, so will the numbers drawn.

(g) **The variance of the histogram for the numbers drawn converges to zero.**

False; this is just (f) in disguise.

(h) **The variance of the average of the draws converges to zero.**

True. The SE for the average is the SD of the box divided by the square root of the numbers drawn. The variance of the average is the square of this. As n goes to infinity, the denominator blows up, making the variance converge to zero.

3. **Match the phrase in the first column of Table 1 to its synonym in the second column. Here, m in the number of units assigned to the treatment group in an experiment with $N > m$ units in the study group.**

(a) The sampling distribution of the treatment group mean	(1) The square root of the variance of the potential outcomes under treatment, divided by the square root of m
(b) The standard error of the mean in the assigned-to-treatment group	(2) The average causal effect of treatment assignment
(c) The estimated standard error of the treatment group mean	(3) the squared s.d. of the sampling distribution of the treatment group mean
(d) The average observed outcome in the assigned-to-treatment group, minus the average observed outcome in the assigned-to-control group	(4) An estimator of the effect of treatment assignment
(e) The sampling variance of the treatment group mean	(5) The standard deviation of observed outcomes in the treatment group, divided by the square root of m
(f) The outcome if every unit were assigned to treatment, minus the outcome if every unit were assigned to control	(6) A histogram showing sample averages in the treatment group, across hypothetical replications of the experiment
(g) The sampling variance of a treatment group "ticket"	(7) The variance of potential outcomes under treatment

Table 1: Match the phrase in the first column to its synonym in the second column.

(a) → (6)

(b) → (1)

(c) → (5) (N.B. remember that the SE for an average of an i.i.d. sample is the standard deviation of the box, here approximated by the sd of the sample, over the square root of the number of draws. See FPP 23, section 2. For the treatment group mean, we are drawing without replacement, so we could multiply by the (square root of the) finite-sample correction factor. For the SE of the difference of means, we would not need to do so, however—we use the "conservative" formula discussed in class.)

(d) → (4)

(e) → (3)

(f) → (2)

(g) → (7)

4. **A large college course has 900 students, broken down into section meetings with 30 students each. The section meetings are led by teaching assistants. On the final, the class average is 63, and the SD is 20. However, in one section the average is only 55. The TA argues this way:**

   "If you took 30 students at random from the class, there is a pretty good chance they would average below 55 on the final. That's what happened to me–chance variation."

   (a) **Formulate the chance procedure the TA describes in terms of a box model. Describe the contents of the box.**

   Here we want to reason from the box (which has 900 tickets in it, one for each student, that records their scores on the exam) to the sample that we observe in this particular classroom. If the sample is random—and that is the thought experiment proposed by the TA—then averaging the scores of 30 students in the class is like averaging the scores of n=30 tickets drawn at random from the box.

   (b) **Fill in the blanks. The null hypothesis says that the average of the box is ___. The alternative hypothesis says that the average of the box is ___.**

   The null hypothesis says that the average of the box is 63. The alternative hypothesis says that the average of the box is less than 63.

   (c) **Is the TAs defense a good one? Answer yes or no, and explain briefly.**

   With n large enough, the probability histogram of the average of draws converges to the normal distribution. So, if we convert the data into standard units, we can use the normal approximation to see whether the outcome we observed, here a section average of 55, can be explained by the luck of the draw. Now, the difference between the score in the TA's section and the average of the tickets in the box (the class average on the exam) is $55 - 63 = -8$. How many standard errors below the average of the box is this difference of -8? To find the standard error, divide the standard deviation of the box by the square root of the number of draws.The standard deviation of the box is 20, the number of draws is 30. Thus,

$$\text{SE} = \frac{20}{\sqrt{30}} \doteq 3.65. \tag{1}$$

   So 55 is $\frac{-8}{3.65} = -2.19$ standard errors below the average of the box.[1]

   Using the normal curve, the percentage of draws that will lie more than 2.19 standard errors away from the average of the box is about 2.8%; you can find this in a standard normal table, such as the one at the back of FPP. (If we ask what percentage of draws will lie more than 2.19 *below* the mean—a *one-tailed test*—the percentage is closer to 1.4%.) So, the TA's defense does not look good: chance variation would only rarely produce such a large difference between the observed average of the sample and the average of the box.

   Should we be using the normal approximation? Well, with 30 draws, the approximation can be quite good, especially if the tickets in the box are normally distributed—and such test scores often are fairly normal.

[1] A finite-population correction factor could also be applied but would make little difference here: $\frac{900-30}{900-1} \doteq .97$, so the corrected SE is about $\sqrt{.97}(3.65) \doteq 3.60$. Then, 55 is about $\frac{-8}{3.60} \doteq -2.22$ standard errors below the average of the box. This only strengthens the case against chance variation.

5. **A geography test was given to a simple random sample of 250 high school students in a certain large school district. One question involved an outline map of Europe, with the countries identified only by number. The students were asked to pick out Great Britain and France. As it turned out, 65.8% could find France, compared to 70.2% for Great Britain. Is the difference statistically significant? Or can this be determined from the information given? Explain.**

   The answer can't be determined from the information given. If we wanted to formulate the sampling procedure as a box model, the box should have four kinds of tickets:$\{1,1\},\{1,0\},\{0,1\},\{0,0\}$. Here, $\{1,1\}$ indicates that the student can find both France and Great Britain, $\{1,0\}$ indicates that the student can find France but not Great Britain, and so on.

   If we want to conduct a statistical test, it should be formulated in terms of this model; for instance, we might want to test the null hypothesis that the proportion who can find France but not Great Britain is equivalent to the proportion who can find Great Britain but not France, against the alternative hypothesis that the number who can find Great Britain but not France is greater than the proportion who can find France but not Great Britain. But we would need more data to conduct this test. For instance, we need information for each sampled student on whether he or she could find Great Britain and/or France, so we know whether the value of each sample ticket is $\{1,1\}$, $\{1,0\}$, $\{0,1\}$, or $\{0,0\}$. Then, we could use the distribution of sample tickets to estimate the distribution of the box, that is, the distribution for all high school students in the large school district.

   Lesson: to conduct a test of statistical significance, you need to be able to construct a box model for the sampling process—and you need the right data to estimate the contents of the box.

6. **Consider an experiment with one treatment group and one control group. There are 6 subjects; 3 subjects are assigned to treatment and 3 are assigned to control. Hypothetical potential outcomes under treatment and control are depicted in Table 1.**

Subject i	$Y_i(1)$	$Y_i(0)$	τ_i
1	5	4	1
2	7	7	0
3	7	3	4
4	6	7	-1
5	6	3	3
6	5	3	2
Average	**6**	**4.5**	**1.5**

Table 1. Hypothetical potential outcomes for the 6 units in this experiment. Here, $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes under treatment and control, while τ_i is the unit causal effect.

You can find R code to produce the analysis at the end below.

(a) **Does the treatment have a positive effect for every unit? On average, does it have a positive effect for all units? Are these quantities parameters or estimators? What is another term for the average τ_i, that is, the number 1.5 found in the bottom-right cell?**

The treatment does not have a positive effect for all units, but on average the effect is positive. These quantities are parameters, they represent the true effects if we could observe each unit in both treatment and control states. The average of the unit causal effects is called the "average causal effect" or the "average treatment effect'."

(b) **How many possible random assignments are there in which 3 units are assigned to treatment and 3 to control?**

The study group is a set of 6 people. We want to know how many ways there are to assign $k = 3$ people to one group. Use the "S choose k" formula for combinations:

$$\binom{S}{k} = \frac{S!}{k!(S-k)!}$$

$$\binom{6}{3} = \frac{6!}{3!(6-3)!} = \frac{6*5*4*3*2*1}{3*2*1*3*2*1} = 20$$

Thus, there are 20 possible random assignments. The table below shows all possible random assignments, with the numbers in the table refering to the index of the units that would receive the treatment in that given assignment (eg. in the first assignment, units 1, 2 and 3 are assigned to the treatment group whereas 4, 5 and 6 would be assigned to the control group).

(c) **Create a table showing, for each possible random assignment, the average response in the treatment group, the average response in the control group, and the difference of means across the treatment and control groups.**

Use Table 2 to find the averages in treatment and control in each possible assignment, then calculate the difference of means. Table 3 shows the result.

(d) **What is the average of the difference of means across the treatment and control groups, across all these possible assignments? What do you conclude about the difference of means, as an estimator for the average causal effect? Give a statistical rationale for your conclusion.**

The average of the difference of means across all possible assignments is 1.5. This is the same as the average causal effect from the schedule of potential outcomes. Since each of the possible assignments is equally likely, we conclude that the difference of means is an unbiased estimator of the average causal effect.

(e) **Calculate the variance of the average responses in the treatment group, across all possible assignments. (That is, compute the mean squared deviation of each treatment group mean from the overall average of the treatment group means, across all assignments).**

Let Y_a^T be the average observed outcome in the treatment group for assignment $a = 1, ..., 20$. Then, let $\overline{Y^T} = \frac{1}{20}\sum_{a=1}^{20} Y_a^T$ be the average of the averages in the treatment group for all possible assignments. We are asked to calculate the variance of the sampling distribution of Y^T using:

$$\text{Var}(Y^T) = \frac{1}{20}\sum_{a=1}^{20}(Y_a^T - \overline{Y^T})^2 \tag{2}$$

	First unit in treatment	Second unit in treatment	Third unit in treatment
1	1	2	3
2	1	2	4
3	1	2	5
4	1	2	6
5	1	3	4
6	1	3	5
7	1	3	6
8	1	4	5
9	1	4	6
10	1	5	6
11	2	3	4
12	2	3	5
13	2	3	6
14	2	4	5
15	2	4	6
16	2	5	6
17	3	4	5
18	3	4	6
19	3	5	6
20	4	5	6

Table 2: All possible assignments to Treatment and Control Groups for six units. The numbers in the table refering to the index of the units that would receive the treatment in that given assignment.

Here, the sampling variance is about 0.13.

(f) **Compare your answer in part (e) to the analytic formula for the sampling variance of the treatment group mean, as presented in class. (Remember that you need a finite-sample correction factor: $\frac{N-m}{N-1}$, where N is the size of the population, and m is the size of the sample).**

The equation for the variance of Y^T using the finite sample correction factor is

$$\text{Var}(Y^T) = \frac{N-m}{N-1} * \frac{\sigma^2}{m}$$

	mean in treatment group	mean in control group	difference in means
assignment 1	6.33	4.33	2.00
assignment 2	6.00	3.00	3.00
assignment 3	6.00	4.33	1.67
assignment 4	5.67	4.33	1.33
assignment 5	6.00	4.33	1.67
assignment 6	6.00	5.67	0.33
assignment 7	5.67	5.67	0.00
assignment 8	5.67	4.33	1.33
assignment 9	5.33	4.33	1.00
assignment 10	5.33	5.67	-0.33
assignment 11	6.67	3.33	3.33
assignment 12	6.67	4.67	2.00
assignment 13	6.33	4.67	1.67
assignment 14	6.33	3.33	3.00
assignment 15	6.00	3.33	2.67
assignment 16	6.00	4.67	1.33
assignment 17	6.33	4.67	1.67
assignment 18	6.00	4.67	1.33
assignment 19	6.00	6.00	0.00
assignment 20	5.67	4.67	1.00

Table 3: Average response in treatment and control group for 20 possible assignments, and the difference in means across these groups.

Here, the finite sample correction factor $\frac{N-m}{N-1} = \frac{6-3}{6-1} = \frac{3}{5}$. Also, $m = 3$, and

$$\begin{aligned}
\sigma^2 &= \text{Var}(Y_i(1)) \\
&= \frac{1}{6}\sum_{i=1}^{6}(Y_i - \overline{Y_i(1)})^2 \\
&= \frac{1}{6}[(5-6)^2 + (7-6)^2 + (7-6)^2 + (6-6)^2 + (6-6)^2 + (5-6)^2] \\
&= \frac{1}{6}[1+1+1+0+0+1] \\
&= \frac{2}{3}. \qquad (3)
\end{aligned}$$

(In the third line, we are just plugging in the six potential outcomes under treatment from the potential outcomes schedule given in the question, and calculating the sum of squared deviations from the average potential outcome under treatment of 6).

Thus, the equation for the sampling variance of the treatment group mean Y^T is

$$\begin{aligned} \text{Var}(Y^T) &= \frac{N-m}{N-1} * \frac{\sigma^2}{m} \\ &= \frac{3}{5}[\frac{\frac{2}{3}}{3}] \\ &= \frac{2}{15} \\ &\doteq 0.13. \end{aligned} \tag{4}$$

The point here is that the analytic formula for the sampling variance of the mean is correct: it gives the same answer as calculating the variance of the means across all 20 possible assignments.

```r
# getting all ways of choosing 3 units out of
# six (numbers in the table refer to the index
# of the element)

treat_assignments <- combn(6, 3)

Y_i1 <- c(5, 7, 7, 6, 6, 5)
Y_i0 <- c(4, 7, 3, 7, 3, 3)


get_means_diffs <- function(assign, y1, y0) {

    mean_treat <- mean(y1[assign])
    mean_control <- mean(y0[-assign])
    diff <- mean_treat - mean_control

    return(c(mean_treat, mean_control, diff))
}


means_by_assign <- apply(treat_assignments, 2,
    FUN = get_means_diffs, y1 = Y_i1, y0 = Y_i0)
means_by_assign <- t(means_by_assign)

colnames(means_by_assign) <- c("mean in treatment group",
    "mean in control group", "difference in means")
rownames(means_by_assign) <- paste("assignment",
    1:20, sep = " ")

means_by_assign
```

```
##               mean in treatment group
## assignment 1                 6.333333
## assignment 2                 6.000000
## assignment 3                 6.000000
## assignment 4                 5.666667
## assignment 5                 6.000000
## assignment 6                 6.000000
## assignment 7                 5.666667
## assignment 8                 5.666667
## assignment 9                 5.333333
## assignment 10                5.333333
## assignment 11                6.666667
## assignment 12                6.666667
## assignment 13                6.333333
## assignment 14                6.333333
## assignment 15                6.000000
## assignment 16                6.000000
## assignment 17                6.333333
## assignment 18                6.000000
## assignment 19                6.000000
## assignment 20                5.666667
##               mean in control group
## assignment 1                4.333333
## assignment 2                3.000000
## assignment 3                4.333333
## assignment 4                4.333333
## assignment 5                4.333333
## assignment 6                5.666667
## assignment 7                5.666667
## assignment 8                4.333333
## assignment 9                4.333333
## assignment 10               5.666667
## assignment 11               3.333333
## assignment 12               4.666667
## assignment 13               4.666667
## assignment 14               3.333333
## assignment 15               3.333333
## assignment 16               4.666667
## assignment 17               4.666667
## assignment 18               4.666667
## assignment 19               6.000000
## assignment 20               4.666667
##               difference in means
```

```
## assignment 1             2.0000000
## assignment 2             3.0000000
## assignment 3             1.6666667
## assignment 4             1.3333333
## assignment 5             1.6666667
## assignment 6             0.3333333
## assignment 7             0.0000000
## assignment 8             1.3333333
## assignment 9             1.0000000
## assignment 10           -0.3333333
## assignment 11            3.3333333
## assignment 12            2.0000000
## assignment 13            1.6666667
## assignment 14            3.0000000
## assignment 15            2.6666667
## assignment 16            1.3333333
## assignment 17            1.6666667
## assignment 18            1.3333333
## assignment 19            0.0000000
## assignment 20            1.0000000

# average of treatment mean across all
# possible assignments
mean(means_by_assign[, 1])

## [1] 6

# average of control mean across all possible
# assignments
mean(means_by_assign[, 2])

## [1] 4.5

# average of the difference of means across
# all possible assignments
mean(means_by_assign[, 3])

## [1] 1.5

# write formula for population variance (R
# uses the n-1 one)
pop_var <- function(x) {
```

```r
    mean((x - mean(x))^2)
}

# variance of the average responses in the
# treatment group, across all possible
# assignments
pop_var(means_by_assign[, 1])

## [1] 0.1333333

# analytic formula for the sampling variance
# of the treatment group mean
sigma2 <- pop_var(Y_i1)
sigma2

## [1] 0.6666667

sampling_var <- function(m, N, sigma2) {
    (N - m)/(N - 1) * sigma2/m
}

sampling_var(m = 3, N = 6, sigma2 = sigma2)

## [1] 0.1333333

# The point here is that the analytic formula
# for the sampling variance of the mean is
# correct: it gives the same answer as
# calculating the variance of the means across
# all 20 possible assignments.
pop_var(means_by_assign[, 1]) == sampling_var(m = 3,
    N = 6, sigma2 = sigma2)

## [1] TRUE
```

7. (R exercise) Using the data from Dunning and Harrison (2010),[2]

   Part A.

   (a) **Expand the function for the difference in means that you wrote in section so that it also calculates and outputs the standard error of the difference in means.**

[2]The data is on bCourses in the folder "Problem Set 2".

```r
diff_se <- function(y, x) {

    # Calculating difference in means
    mean1 <- mean(y[x == 1], na.rm = T)
    mean0 <- mean(y[x == 0], na.rm = T)
    diff <- mean1 - mean0

    # Calculating SE of the difference
    N1 <- length(na.omit(y[x == 1]))
    N0 <- length(na.omit(y[x == 0]))
    var1 <- var(y[x == 1], na.rm = T)
    var0 <- var(y[x == 0], na.rm = T)
    varN1 <- var1/N1
    varN0 <- var0/N0
    se.diff <- sqrt(varN1 + varN0)

    # Preparing output
    res <- c(mean1, mean0, diff, se.diff, (N1 +
        N0))
    names(res) <- c("Mean 1", "Mean 0", "Difference",
        "SE Diff", "N")

    return(c(res))
}

# let's code an arbitrary example to see if it
# works
diff_se(y = c(rep(2, 100), rep(4, 100)), x = c(rep(0,
    100), rep(1, 200)))
```

```
##     Mean 1     Mean 0 Difference    SE Diff
##          4          2          2          0
##          N
##        200
```

```r
# QUESTION: Why is the SE of the difference 0
# here?
```

(b) **Replicate tables 3 and 4 in the paper, i.e. write code that produces tables that show the results in the paper. (Note: do not use the t-test function in R. Instead, use the function you wrote for part 1. You just need the means, differences of means, and standard errors here; do not worry about p-values or hypothesis testing).**

```r
# Let's give the data a look
names(data)
```

```
##  [1] "participantid"
##  [2] "nom"
##  [3] "treat_assign"
##  [4] "actor"
##  [5] "name_pol"
##  [6] "ethnic_pol"
##  [7] "interviewer"
##  [8] "date_interview"
##  [9] "X_2nd_treatment"
## [10] "female"
## [11] "naissance"
## [12] "classe"
## [13] "education"
## [14] "annees_bamako"
## [15] "vecu_ailleurs"
## [16] "inscrite"
## [17] "daily_language"
## [18] "ethnic_subject"
## [19] "global_eval"
## [20] "global_eval01"
## [21] "vote_prefer"
## [22] "vote_prefer01"
## [23] "aimable"
## [24] "aimable01"
## [25] "intelligent"
## [26] "intelligent01"
## [27] "digne_confiance"
## [28] "digne_confiance01"
## [29] "competent"
## [30] "competent01"
## [31] "defis"
## [32] "defis01"
## [33] "impressione"
## [34] "impressione01"
## [35] "idees"
## [36] "idees01"
## [37] "motivations"
## [38] "motivations01"
## [39] "boulot"
## [40] "boulot01"
## [41] "defendrait"
```

```
## [42] "defendrait01"
## [43] "ethnie_candidat"
## [44] "ethnie_autre"
## [45] "ethnic_correct"
## [46] "ethnie_match"
## [47] "cousin_subject"
## [48] "cousin_match"
## [49] "cousin_assign"
## [50] "self_assign"
## [51] "self_ass_alt"
## [52] "nom_faites_attention"
## [53] "voter_memenom"
## [54] "voter_cousin"
## [55] "penser_autresdiscours"
## [56] "parti_politique"
## [57] "quel_parti"
## [58] "vote_pres2007"
## [59] "qui_pres2007"
## [60] "vote_leg2007"
## [61] "amis_nom_premier"
## [62] "connaissance_premier"
## [63] "marier_autreethnie"

# Table 3 - Descriptive statistics on response variables

# Variables
# Global evaluation of candidate
# Global evaluation of speech
# Is likeable
# Is intelligent
# Is competent
# Is impressive
# Is trustworthy
# Would do a good job in office
# Would defend others and fight for his ideals
# Has good motivations for running
# Would successfully face challenges of office
# Has good ideas

# For each variable, we want the range, the mean and the SD

# We want to use variables from 19 to 42

cc_desc <- data[,19:42]
```

```r
# we will use apply to run the function over each column (margin=2) for all the
# variables in just one line

cc_mean <- apply(cc_desc, 2, FUN=mean, na.rm=TRUE)
cc_sd <- apply(cc_desc, 2, FUN=sd, na.rm=TRUE)
cc_range <- apply(cc_desc, 2, FUN=range, na.rm=TRUE)

names(cc_desc)
```

```
##  [1] "global_eval"       "global_eval01"
##  [3] "vote_prefer"       "vote_prefer01"
##  [5] "aimable"           "aimable01"
##  [7] "intelligent"       "intelligent01"
##  [9] "digne_confiance"   "digne_confiance01"
## [11] "competent"         "competent01"
## [13] "defis"             "defis01"
## [15] "impressione"       "impressione01"
## [17] "idees"             "idees01"
## [19] "motivations"       "motivations01"
## [21] "boulot"            "boulot01"
## [23] "defendrait"        "defendrait01"
```

```r
cc_range # we want to flip this so that we have to columns rather than
```

```
##      global_eval global_eval01 vote_prefer
## [1,]           1             0           1
## [2,]           7             1           7
##      vote_prefer01 aimable aimable01
## [1,]             0       1         0
## [2,]             1       5         1
##      intelligent intelligent01
## [1,]           1             0
## [2,]           5             1
##      digne_confiance digne_confiance01
## [1,]               1                 0
## [2,]               5                 1
##      competent competent01 defis defis01
## [1,]         1           0     1       0
## [2,]         5           1     7       1
##      impressione impressione01 idees idees01
## [1,]           1             0     1       0
## [2,]           7             1     7       1
##      motivations motivations01 boulot
## [1,]           1             0      1
```

```
## [2,]           7             1      7
##      boulot01 defendrait defendrait01
## [1,]        0          1            0
## [2,]        1          7            1

# two rows
cc_range <- aperm(cc_range)
head(cc_range)

##                [,1] [,2]
## global_eval       1    7
## global_eval01     0    1
## vote_prefer       1    7
## vote_prefer01     0    1
## aimable           1    5
## aimable01         0    1

# when I used apply, I got the output for all the variables, which include
# the same variable using both the 1-7 and 0-1 scale. I know want to keep
# every other column (only the ones in the 1-7 scale)

seq(1,length(names(cc_desc)), by=2) # this sequence give me the indices

##  [1]  1  3  5  7  9 11 13 15 17 19 21 23

# for the variables I want to keep

table_3 <- as.data.frame(cbind(
  # for the first to columns I used the indexes above so that I only keep
  # the descriptive stats of the 1-7 scale
  cc_range[seq(1,length(names(cc_desc)), by=2),],
  cc_mean[seq(1,length(names(cc_desc)), by=2)],
  cc_sd[seq(1,length(names(cc_desc)), by=2)],
  # and for the last two columns I change the sequence so that I get the 0-1 sca
  cc_mean[seq(2,length(names(cc_desc)), by=2)],
  cc_sd[seq(2,length(names(cc_desc)), by=2)]))

names(table_3) <- c("Min value", "Max value", "Mean", "SD",
                    "Mean in 0-1 scale", "SD in 0-1 scale")

table_3

##                 Min value Max value     Mean
## global_eval             1         7 6.287621
## vote_prefer             1         7 4.530340
```

```
## aimable                  1         5 4.484812
## intelligent              1         5 2.904010
## digne_confiance          1         5 2.574029
## competent                1         5 2.719854
## defis                    1         7 3.996350
## impressione              1         7 4.257631
## idees                    1         7 6.012180
## motivations              1         7 6.126521
## boulot                   1         7 3.488457
## defendrait               1         7 2.991995
##                        SD Mean in 0-1 scale
## global_eval     1.2149503         0.8812702
## vote_prefer     1.7307320         0.5883900
## aimable         0.6076961         0.8712029
## intelligent     0.9522778         0.4760024
## digne_confiance 1.0831103         0.3935073
## competent       0.9637113         0.4299635
## defis           1.3530266         0.4993917
## impressione     1.6944609         0.5429385
## idees           1.4381044         0.8353634
## motivations     1.3885042         0.8544201
## boulot          1.7776801         0.4147428
## defendrait      1.8022448         0.3319992
##                 SD in 0-1 scale
## global_eval           0.2024917
## vote_prefer           0.2884553
## aimable               0.1519240
## intelligent           0.2380695
## digne_confiance       0.2707776
## competent             0.2409278
## defis                 0.2255044
## impressione           0.2824102
## idees                 0.2396841
## motivations           0.2314174
## boulot                0.2962800
## defendrait            0.3003741

# I can even use some rounding so that I don't get so many decimal places
round(table_3, digits=3)

##                 Min value Max value  Mean
## global_eval             1         7 6.288
## vote_prefer             1         7 4.530
## aimable                 1         5 4.485
```

```
## intelligent             1         5 2.904
## digne_confiance         1         5 2.574
## competent               1         5 2.720
## defis                   1         7 3.996
## impressione             1         7 4.258
## idees                   1         7 6.012
## motivations             1         7 6.127
## boulot                  1         7 3.488
## defendrait              1         7 2.992
##                    SD Mean in 0-1 scale
## global_eval     1.215             0.881
## vote_prefer     1.731             0.588
## aimable         0.608             0.871
## intelligent     0.952             0.476
## digne_confiance 1.083             0.394
## competent       0.964             0.430
## defis           1.353             0.499
## impressione     1.694             0.543
## idees           1.438             0.835
## motivations     1.389             0.854
## boulot          1.778             0.415
## defendrait      1.802             0.332
##                 SD in 0-1 scale
## global_eval               0.202
## vote_prefer               0.288
## aimable                   0.152
## intelligent               0.238
## digne_confiance           0.271
## competent                 0.241
## defis                     0.226
## impressione               0.282
## idees                     0.240
## motivations               0.231
## boulot                    0.296
## defendrait                0.300

# Table 4 - Average candidate evaluations by treatment assignment

# treat_assign takes on a value 1 through 6 and denotes the treatment condition
# which the respondent was assigned, as follows:
# 1 -- Same ethnicity, joking cousin
# 2 -- Same ethnicity, not joking cousin
# 3 -- Different ethnicity, joking cousin
# 4 -- Different ethnicity, not joking cousin
```

```r
# 5 -- No last name given for candidate
# 6 -- Candidate and subject have same last name (and thus ethnicity)

# we can write a small loop to get mean and sd by treatment

mean_bytreat <- NA
sd_bytreat <- NA

for (i in 1:6){
    mean_bytreat[i] <- mean(data$vote_prefer[data$treat_assign==i])
    sd_bytreat[i] <- sd(data$vote_prefer[data$treat_assign==i])/
                    sqrt(length(data$vote_prefer[data$treat_assign==i]))
    }

table_4 <- rbind(mean_bytreat, sd_bytreat)
colnames(table_4)  <- c("Same ethnicity, joking cousin",
                        "Same ethnicity, not joking cousin",
                        "Different ethnicity, joking cousin",
                        "Different ethnicity, not joking cousin",
                        "No last name given for candidate",
                        "Candidate and subject have same last name (and thus eth

rownames(table_4) <- c("mean", "sd")

table_4
```

```
##      Same ethnicity, joking cousin
## mean                     5.0514706
## sd                       0.1495527
##      Same ethnicity, not joking cousin
## mean                         4.5655738
## sd                           0.1552878
##      Different ethnicity, joking cousin
## mean                          4.4435484
## sd                            0.1652718
##      Different ethnicity, not joking cousin
## mean                                3.9605263
## sd                                  0.1296433
##      No last name given for candidate
## mean                         4.325758
## sd                           0.119997
##      Candidate and subject have same last name (and thus ethnicity)
## mean                                                       4.8417722
## sd                                                         0.1474123
```

```r
# Now we are missing the part of the table that compares individuals from the sa
# ethnicity, joking cousins vs not.

# (For this, you could have also used the function you wrote for 11a)

# keeping only the data with the treatment groups we care about
diff_12 <- with(data[data$treat_assign==1 | data$treat_assign==2,],
     # and running the t-test. treat_assign==1 turns treat_assign in a dummy
     # with treat_assign equal to 1 when it is 1 and 0 if the original variable
     # is 2
     diff_se(vote_prefer, treat_assign==1)
     )

diff_34 <- with(data[data$treat_assign==3 | data$treat_assign==4,],
     # and running the t-test. treat_assign==1 turns treat_assign in a dummy
     # with treat_assign equal to 1 when it is 1 and 0 if the original variable
     # is 2
     diff_se(vote_prefer, treat_assign==3)
     )

diff_13 <- with(data[data$treat_assign==1 | data$treat_assign==3,],
     # and running the t-test. treat_assign==1 turns treat_assign in a dummy
     # with treat_assign equal to 1 when it is 1 and 0 if the original variable
     # is 2
     diff_se(vote_prefer, treat_assign==1)
     )

diff_24 <- with(data[data$treat_assign==2 | data$treat_assign==4,],
     # and running the t-test. treat_assign==1 turns treat_assign in a dummy
     # with treat_assign equal to 1 when it is 1 and 0 if the original variable
     # is 2
     diff_se(vote_prefer, treat_assign==2)
     )


diff_12
```

```
##      Mean 1      Mean 0  Difference
##   5.0514706   4.5655738   0.4858968
##     SE Diff           N
##   0.2155929 258.0000000
```

```r
diff_34
```

```
##      Mean 1      Mean 0  Difference
```

```
##   4.4435484   3.9605263   0.4830221
##     SE Diff           N
##   0.2100527 276.0000000

diff_13

##      Mean 1      Mean 0  Difference
##   5.0514706   4.4435484   0.6079222
##     SE Diff           N
##   0.2228918 260.0000000

diff_24

##      Mean 1      Mean 0  Difference
##   4.5655738   3.9605263   0.6050475
##     SE Diff           N
##   0.2022911 274.0000000

# and now we extract the elements we want and bidn them into a table
diff_means <- cbind(diff_12[3:4], diff_34[3:4], diff_13[3:4], diff_24[3:4])

diff_means

##                 [,1]      [,2]      [,3]
## Difference 0.4858968 0.4830221 0.6079222
## SE Diff    0.2155929 0.2100527 0.2228918
##                 [,4]
## Difference 0.6050475
## SE Diff    0.2022911

rownames(diff_means) <- c("Difference in means", "SE")
colnames(diff_means) <- c("Same ethnicity, joking vs not",
                          "Different ethnicity, joking vs not",
                          "Joking, same vs. different ethnicity",
                          "Not joking, same vs. different ethnicity")

round(diff_means, 3)

##                     Same ethnicity, joking vs not
## Difference in means                         0.486
## SE                                          0.216
##                     Different ethnicity, joking vs not
## Difference in means                              0.483
## SE                                               0.210
##                     Joking, same vs. different ethnicity
## Difference in means                                0.608
```

```
## SE                                                    0.223
##                       Not joking, same vs. different ethnicity
## Difference in means                                     0.605
## SE                                                      0.202
```

Part B.

(a) **Carefully define the sharp (strict) null hypothesis involving the "coethnic cousin" treatment condition and the "non-coethnic, non-cousin" control condition. (Be clear about the population for which the null hypothesis is defined).**

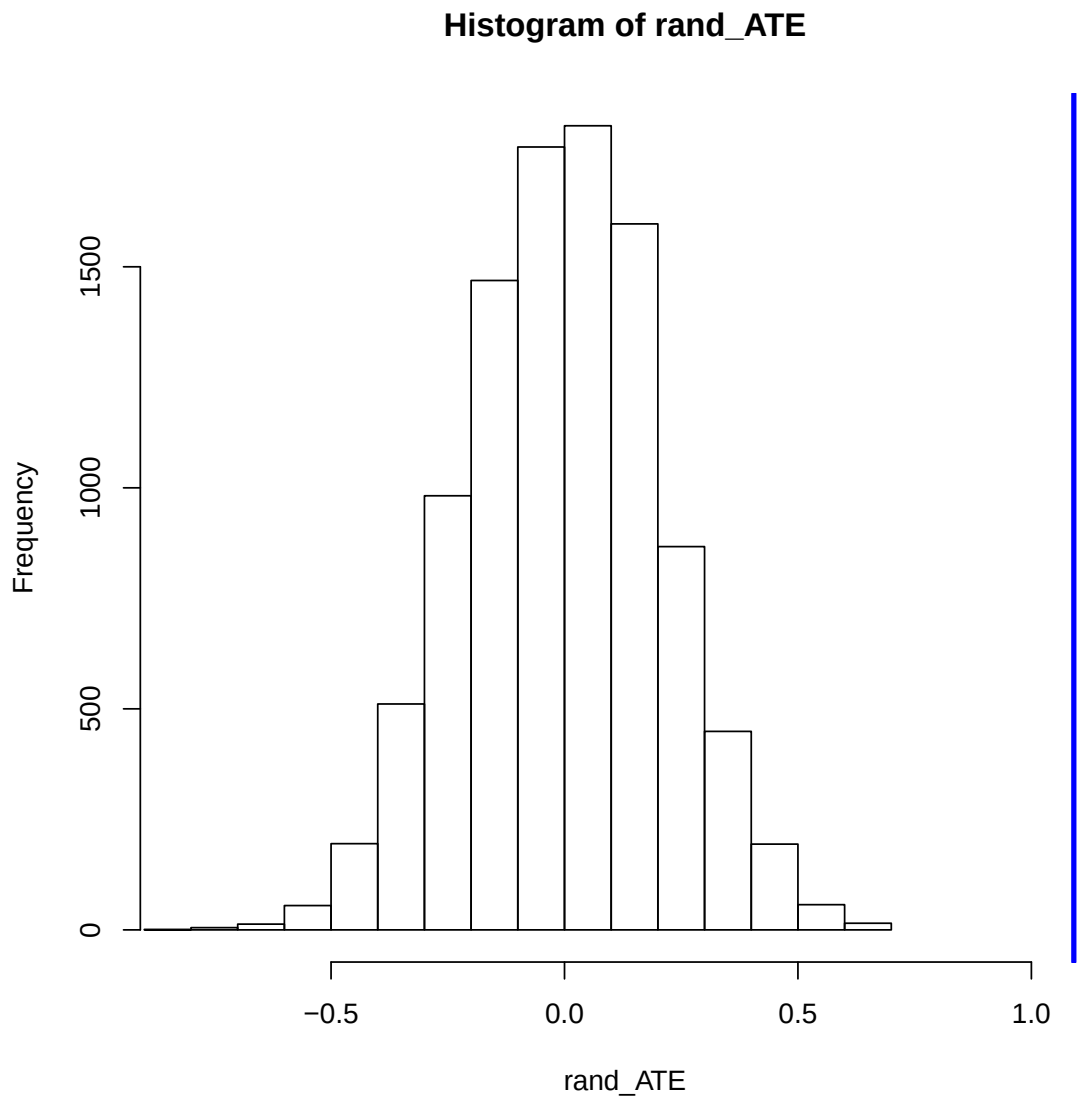
The sharp null states there is no individual effect for any of the units in the population, which here refers to all the individuals in the study who got either the "coethnic cousin" treatment condition or the "non-coethnic, non-cousin" control condition.

(b) **Use randomization inference to test the sharp null hypothesis.**

Note that the problem did not mention which was the outcome variable so this problem could have been completed with other outcomes.

```
# treat_assign takes on a value 1 through 6
# and denotes the treatment condition to which
# the respondent was assigned, as follows: 1
# -- Same ethnicity, joking cousin 2 -- Same
# ethnicity, not joking cousin 3 -- Different
# ethnicity, joking cousin 4 -- Different
# ethnicity, not joking cousin 5 -- No last
# name given for candidate 6 -- Candidate and
# subject have same last name (and thus
# ethnicity)

# (a) We want to compare 'coethnic cousin' to
# 'non-coethnic, non-cousin', so treatment 1
# vs treatment 4.

data <- data[data$treat_assign == 1 | data$treat_assign ==
    4, ]  # keep only observations for which
# treatment is 1 or 4.

# now redefine variable such that it is a
# dummy equal to 1 when treat_assign==1
data$treat_assign <- ifelse(data$treat_assign ==
    1, 1, 0)

# calculate the observed difference of means
ATE <- mean(data$vote_prefer[data$treat_assign ==
```

```
    1]) - mean(data$vote_prefer[data$treat_assign ==
    0])
ATE

## [1] 1.090944

# generate matrix with 10,000 possible
# assignment vectors
randomizations <- matrix(NA, nrow(data), 10000)
for (i in 1:10000) {
    randomizations[, i] <- sample(data$treat_assign,
        length(data$treat_assign), replace = F)
}
randomizations <- unique(randomizations)

# and now for each we compute the ATE we would
# have observed (assuming no unit treatment
# effect in order to obtain a matrix of
# potential outcomes)
rand_ATE <- NA
for (i in 1:10000) {
    # note we index here using the columns in the
    # randomizations matrix
    rand_ATE[i] <- mean(data$vote_prefer[randomizations[,
        i] == 1]) - mean(data$vote_prefer[randomizations[,
        i] == 0])
}

# let's plot the distrbution of randomization
# ATEs and add a line for the observed ATE.
hist(rand_ATE, xlim = c(min(rand_ATE, ATE), max(rand_ATE,
    ATE)))
abline(v = ATE, lwd = 3, col = "blue")

# and now calculate a two-tailed p-value
sum(abs(rand_ATE) >= ATE)/length(rand_ATE)

## [1] 0

# and the p-value is almost zero.
round(sum(abs(rand_ATE) >= ATE)/length(rand_ATE),
    4)

## [1] 0
```



(c) **Carefully define the weak null hypothesis involving the "coethnic cousin" treatment condition and the "non-coethnic, non-cousin" control condition. (Be clear about the population for which the null hypothesis is defined).**

For those assigned to to the "coethnic cousin" treatment or the "non-coethnic, non-cousin" control condition (our population here), the weak null states that there is no average effect– that is, the average of the potential outcomes under treatment equals the average of the potential outcomes under control

(d) **Building on the code you have used in section, write a *t*-test function in R that takes the treatment and the outcome data and returns the difference of means, its standard error and the p-value. Use this function to test the weak null hypothesis.**

```r
# T-test function with SEs and p-value
ttest <- function(y, x) {

    # Calculating difference in means
    mean1 <- mean(y[x == 1], na.rm = T)
    mean0 <- mean(y[x == 0], na.rm = T)
    diff <- mean1 - mean0

    # Calculating SE of the difference
    N1 <- length(na.omit(y[x == 1]))
    N0 <- length(na.omit(y[x == 0]))
    var1 <- var(y[x == 1], na.rm = T)
    var0 <- var(y[x == 0], na.rm = T)
    varN1 <- var1/N1
    varN0 <- var0/N0
    se.diff <- sqrt(varN1 + varN0)

    # T-statistic
    t <- diff/se.diff

    # Degrees of freedom
    df.num <- ((varN1 + varN0)^2)
    df.den <- (varN1^2)/(N1 - 1) + (varN0^2)/(N0 -
        1)
    df <- df.num/df.den

    # P-value
    if (t >= 0) {
        p <- pt(t, df, lower.tail = F) + pt(-t,
            df, lower.tail = T)
    }
    if (t < 0) {
        p <- pt(t, df, lower.tail = T) + pt(-t,
            df, lower.tail = F)
    }

    # Preparing output
    res <- c(mean1, mean0, diff, se.diff, t, (N1 +
        N0), df, p)
    names(res) <- c("Mean 1", "Mean 0", "Difference",
        "SE Diff", "t-stat", "N", "df", "p-value")

    return(c(res))
```

```
}

# test using global_eval as outcome
ttest(y = data$vote_prefer, x = (data$treat_assign ==
    1))

##       Mean 1       Mean 0   Difference
## 5.051471e+00 3.960526e+00 1.090944e+00
##      SE Diff       t-stat            N
## 1.979227e-01 5.511972e+00 2.880000e+02
##           df      p-value
## 2.751947e+02 8.163331e-08

# we could use some rounding
round(ttest(y = data$vote_prefer, x = (data$treat_assign ==
    1)), 4)

##     Mean 1     Mean 0 Difference    SE Diff
##     5.0515     3.9605     1.0909     0.1979
##     t-stat          N         df    p-value
##     5.5120   288.0000   275.1947     0.0000
```

(e) **Compare your *p*-values from the (b) and (d). In which case (or neither, or both) would you reject the null hypothesis? If your *p*-values are very different, why are they different? If they are very similar, why are they similar?**

Here, the p-values are very similar and we would reject the null hypothesis at standard significance levels for both. In general, the p-value from a randomization test will converge to the p-value from a t-test as the number of units gets large; with 136 units, the study group is fairly large here.

8. **Is the standard deviation of the sample an unbiased estimator of the standard deviation in the population? Does your answer depend on whether you are sampling with or without replacement? Use R to write simulations to answer these questions. Note that R uses (n-1) for the denominator of sd( ).**

For this problem, we want see whether the sd of the sample is an unbiased estimator of the sd of the population. To do this, we will want to compare the sampling distribution of the standard deviation of the sample (with and without replacement) to the true standard deviation of the population.

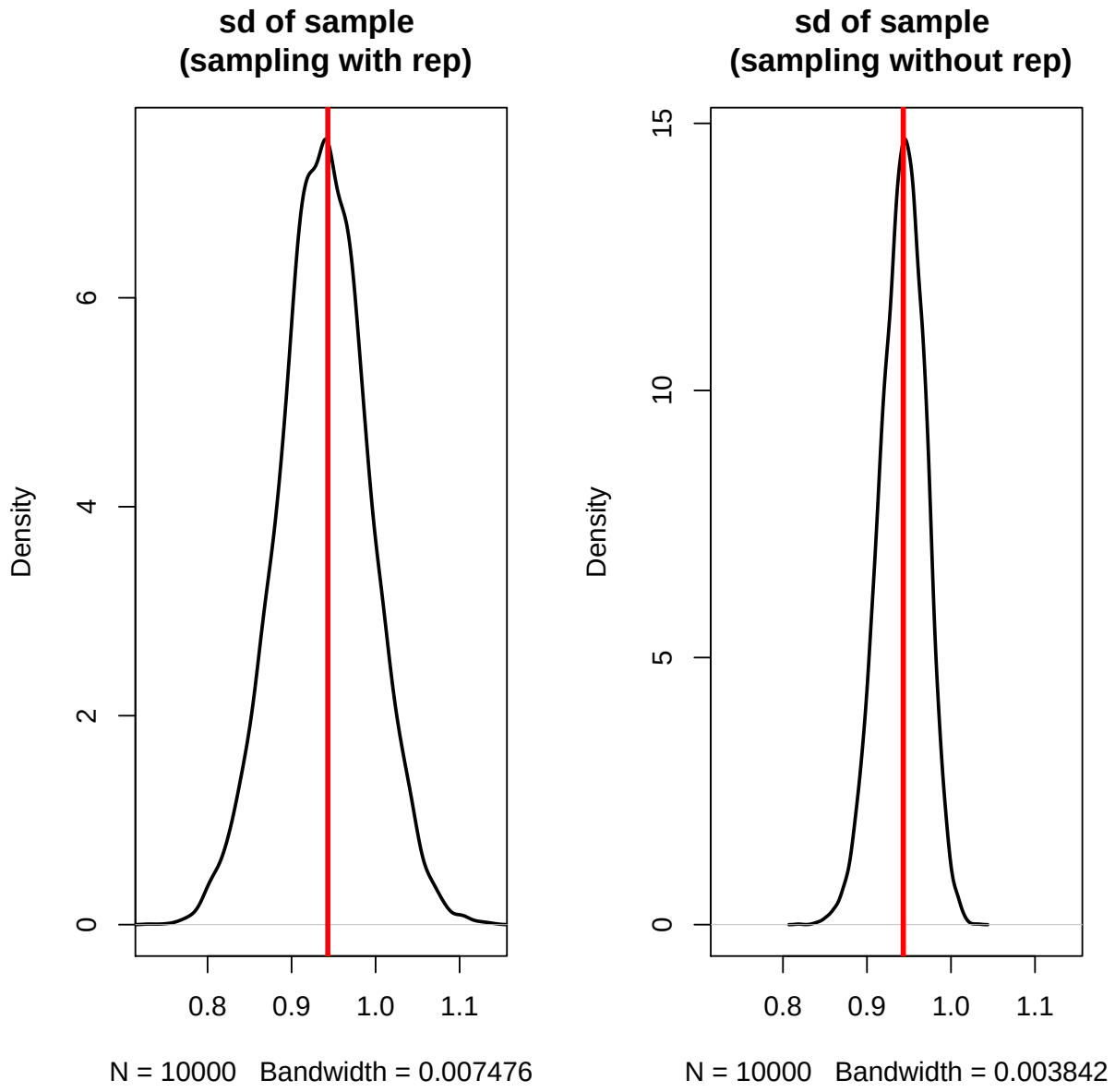
The first step, is to create a population and calculate its standard deviation.

```
set.seed(94705)
# we use rnorm to sample randomly from the
# normal distribution with mean = 0 and sd =
# 1.
population <- rnorm(200, mean = 0, sd = 1)
```

```r
pop.sd <- sqrt(mean((population - mean(population))^2))
pop.sd

## [1] 0.9431266
```

Now we need to generate the sampling distribution of the sample sd. For that, we use a loop in which we sample with and without replacement. We will take a sample of 150 units.

```r
sd.with.rep <- NA
sd.without.rep <- NA

for (i in 1:10000) {

    # generating samples
    with <- sample(population, 150, replace = TRUE)
    without <- sample(population, 150, replace = FALSE)

    # calculating the sd ans storing
    sd.with.rep[i] <- sqrt(mean((with - mean(with))^2))
    sd.without.rep[i] <- sqrt(mean((without -
        mean(without))^2))

}

# and plot
par(mfrow = c(1, 2))  #puts two plots side by side
plot(density(sd.with.rep), main = "sd of sample \n (sampling with rep)",
    lwd = 2, xlim = c(min(c(sd.with.rep, sd.without.rep)),
        max(c(sd.with.rep, sd.without.rep))))
abline(v = pop.sd, col = "red", lwd = 3)
plot(density(sd.without.rep), main = "sd of sample \n (sampling without rep)",
    lwd = 2, xlim = c(min(c(sd.with.rep, sd.without.rep)),
        max(c(sd.with.rep, sd.without.rep))))
abline(v = pop.sd, col = "red", lwd = 3)
```



The plots show that the standard deviation of the sample is an unbiased estimator of the sd for the population both sampling with and without replacement: both sampling distributions are centered around the true value of the population standard deviation.

9. **For this question, you will compare the true standard error of $\widehat{\text{ATE}}$ to the "conservative" standard error.**

   (a) First, consider the following R code:

```
set.seed(1234567)
N <- 60
m <- 30
```

```r
y0 <- rnorm(N, 2, 3)
y1 <- y0 + rnorm(N, 1, 2)
# y0 and y1 are potential outcomes; the ATE is
# about 1
data <- data.frame(y0, y1)
```

(b) **Suppose that `m` units are assigned at random to treatment, with `N-m` assigned to control. Is the difference between the average `y1` in the treatment group and the average `y0` in the control group a random variable?**

Yes, it is a random variable. There is a well-defined chance procedure that assigns units to either group.

(c) **What is the true standard error of this difference of means? What is the "conservative" standard error? (Note: in both cases we are asking about standard errors defined by parameters—not sample quantities).**

The true standard error of the difference of means is:

$$\text{SE}(\hat{\text{ATE}}) = \sqrt{\text{Var}(Y^T) + \text{Var}(Y^C) - 2 * \text{Cov}(Y^T, Y^C)}$$

where

$$\text{Var}(Y^T) = \frac{N-m}{N-1}\frac{\text{Var}(Y_i(1))}{m}$$

$$\text{Var}(Y^C) = \frac{N-(N-m)}{N-1}\frac{\text{Var}(Y_i(0))}{N-m}$$

$$\text{Cov}(Y^T, Y^C) = -\frac{\text{Cov}(Y_i(1), Y_i(0))}{N-1}$$

.

And the conservative standard error:

$$\text{SE}(\hat{\text{ATE}}) = \sqrt{\frac{\text{Var}(Y_i(1))}{N-m} + \frac{\text{Var}(Y_i(0))}{m}}$$

(d) **Now, complete the code above to build a simulation in which there are 10,000 replicates. In each replicate, `m` of the units are assigned at random to treatment. Save the conservative $\widehat{\text{SE}}(\widehat{\text{ATE}})$ for each replicate. Plot the distribution of the conservative $\widehat{\text{SE}}(\widehat{\text{ATE}})$s across the 10,000 replicates. Add a vertical line to your plot at the value of the true standard error.**

```r
set.seed(1234567)
# simulating potential outcomes
N <- 60
m <- 30
```

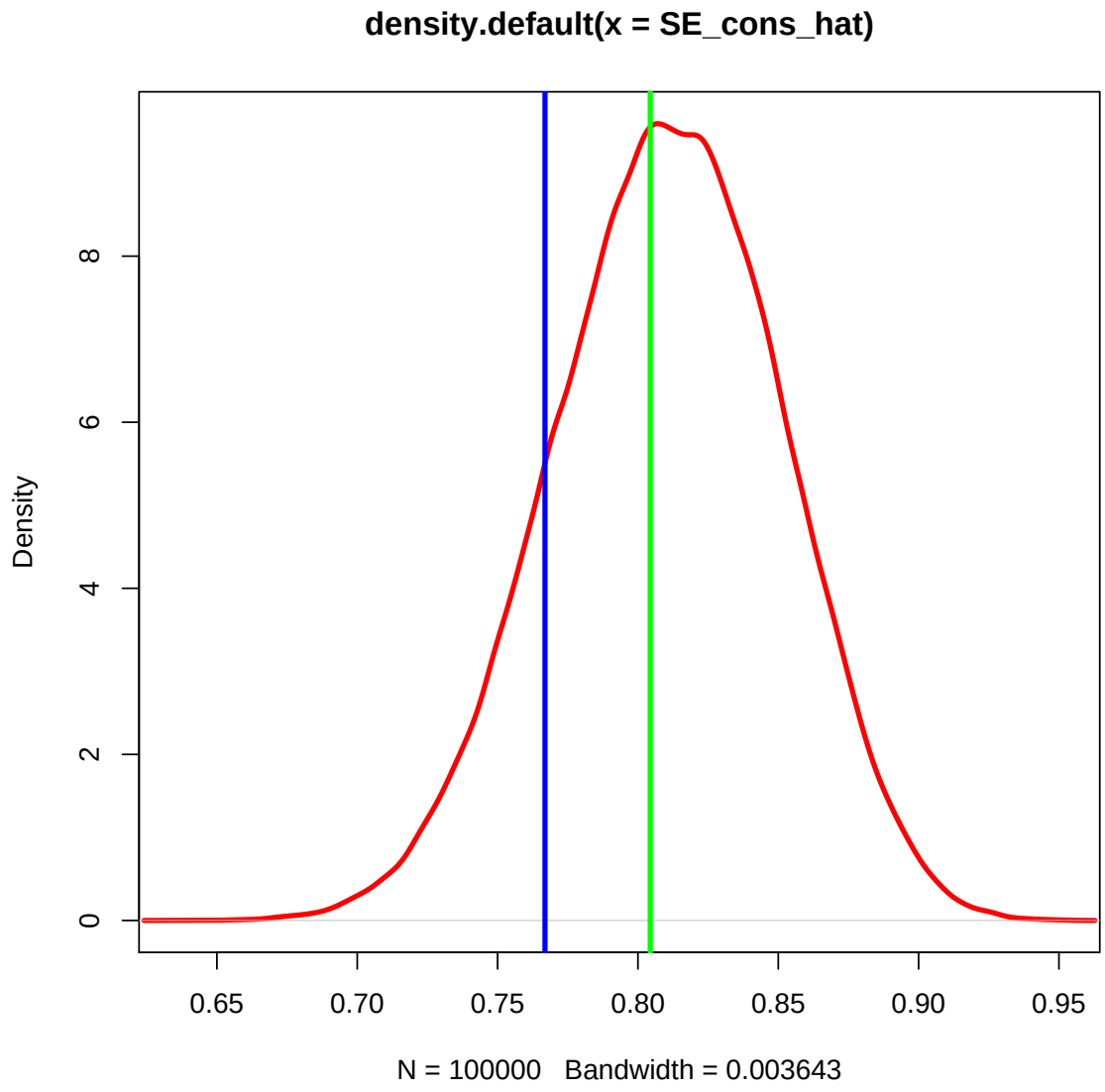
```r
y0 <- rnorm(N, 2, 3)
y1 <- y0 + rnorm(N, 1, 2)  # ATE about 1
data <- data.frame(y0, y1)

replicates <- 1e+05
treat <- c(rep(0, m), rep(1, m))

correc <- (N - m)/(N - 1)

varN <- function(x) {
    mean((x - mean(x))^2)
}

var.yT <- (N - m)/(N - 1) * varN(y1)/m
var.yC <- (N - (N - m))/(N - 1) * varN(y0)/(N -
    m)
cov.yT.yC <- -cov(y1, y0)/(N - 1)

var.yT.yC <- var.yT + var.yC - 2 * cov.yT.yC
SE <- sqrt(var.yT.yC)

SE_cons <- sqrt(varN(y1)/m + varN(y0)/(N - m))

# placeholders
SE_cons_hat <- NA

for (i in 1:replicates) {

    data$t <- sample(treat, length(treat), replace = F)
    data$y_obs <- ifelse(data$t == 1, data$y1,
        data$y0)

    SE_cons_hat[i] <- sqrt(var(data$y_obs[data$t ==
        1])/m + var(data$y_obs[data$t == 0])/(N -
        m))

}

mean(SE_cons_hat)  #statistic
```

```
## [1] 0.8101281
```

```r
SE_cons  # parameter
```

```
## [1] 0.8043778
```

```r
SE  #parameter

## [1] 0.7668868

plot(density(SE_cons_hat), lwd = 3, col = "red",
    xlim = c(min(SE, min(SE_cons_hat)), max(SE_cons_hat)))
abline(v = SE, lwd = 3, col = "blue")
abline(v = SE_cons, lwd = 3, col = "green")
```



(e) **Referring to your plot from part (d), explain why the estimated standard error is called "conservative"?**

The estimated standard error is called conservative because it overestimates the true standard error. This can be seen in the plot in that (1) the mean of the sampling distribution (green

line) is to the right of the parameter value of the true standard error (blue) and (2) most of the sampling distribution of the conservative standard error is to the right of the true SE($\widehat{\text{ATE}}$). If we are conducting hypothesis tests, a larger estimated standard error may lead us to fail to reject the null hypothesis–which is why the larger standard error is "conservative".

Note though that the conservative estimation approach tends to overestimate the true sampling variance *on average*, but the estimation of the conservative standard error is also subject to sampling variability. A given estimate of the sampling variance using the conservative formula may still be smaller than the true sampling variance.

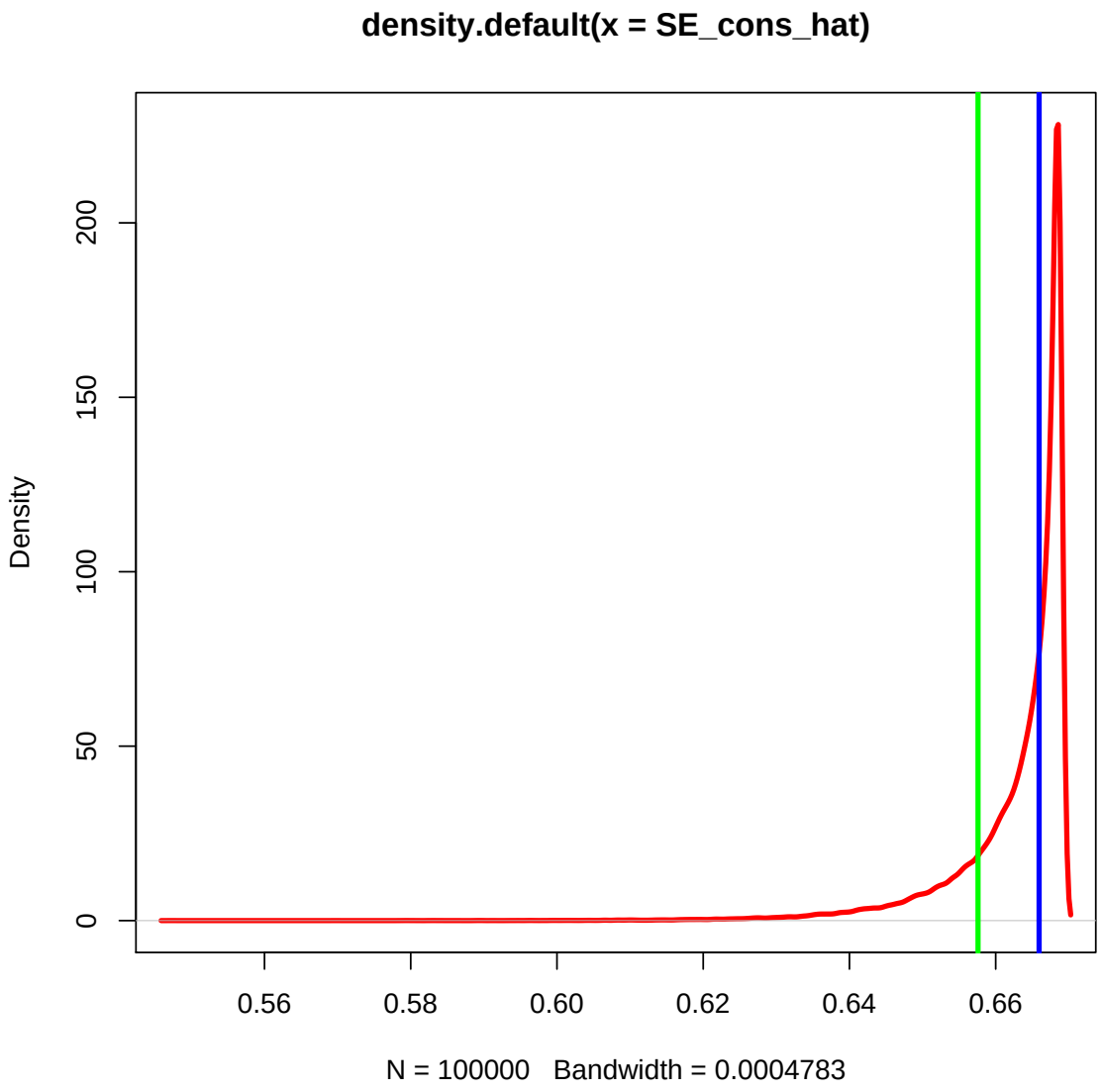
Note, however, that the conservative standard error given in the answer to part (c)–which is defined in terms of population parameters, not sample estimators–is always at least as large as the true standard error of the difference of means also given in the answer to part (c). In this sense, this standard error is indeed "conservative."

(f) **Under what conditions would the true standard error equal the conservative standard error in part (a)? Modify the code excerpted above so that this equality holds.**

We can modify the DGP so that y1 = y0 + a constant. See Dunning (2012) appendix 6.1 on this point.

```
set.seed(1234567)
# simulating potential outcomes
N <- 60
m <- 30
y0 <- rnorm(N, 2, 3)
y1 <- y0 + 3  # ATE 3 for all units
data <- data.frame(y0, y1)


replicates <- 1e+05
treat <- c(rep(0, m), rep(1, (N - m)))

correc <- (N - m)/(N - 1)

varN <- function(x) {
    mean((x - mean(x))^2)
}

var.yT <- (N - m)/(N - 1) * varN(y1)/m
var.yC <- (N - (N - m))/(N - 1) * varN(y0)/(N -
    m)
cov.yT.yC <- -cov(y1, y0)/(N - 1)

var.yT.yC <- var.yT + var.yC - 2 * cov.yT.yC
SE <- sqrt(var.yT.yC)
```

```r
SE_cons <- sqrt(varN(y1)/m + varN(y0)/(N - m))

# placeholders
SE_cons_hat <- NA

for (i in 1:replicates) {

    data$t <- sample(treat, length(treat), replace = F)
    data$y_obs <- ifelse(data$t == 1, data$y1,
        data$y0)

    SE_cons_hat[i] <- sqrt(var(data$y_obs[data$t ==
        1])/m + var(data$y_obs[data$t == 0])/(N -
        m))

}

mean(SE_cons_hat)  #statistic

## [1] 0.6630604

SE_cons  # parameter

## [1] 0.6575861

SE  #parameter

## [1] 0.6659395

plot(density(SE_cons_hat), lwd = 3, col = "red",
    xlim = c(min(SE, min(SE_cons_hat)), max(SE_cons_hat)))
abline(v = SE, lwd = 3, col = "blue")
abline(v = SE_cons, lwd = 3, col = "green")
```


density.default(x = SE_cons_hat)
Density
0
50
100
150
200
0.56
0.58
0.60
0.62
0.64
0.66
N = 100000 Bandwidth = 0.0004783