Methods Subfield Questions

Pol211, UC Davis

Hanno Hilbig

Problem Set 1

In this problem set, we will analyze the relationship between various demographic traits and pro-feminist voting behavior among circuit court judges. In a paper, Adam N. Glynn and Maya Sen argue that having a female child causes circuit court judges to make more pro-feminist decisions. The paper can be found at:

> Glynn, Adam N., and Maya Sen. (2015). "Identifying Judicial Empathy: Does Having Daughters Cause Judges to Rule for Women's Issues?." *American Journal of Political Science* Vol. 59, No. 1, pp. 37–54.

The dataset `dbj.csv` contains the following variables about individual judges:

Name	Description
`name`	The judge's name
`child`	The number of children each judge has.
`circuit.1`	Which federal circuit the judge serves in.
`girls`	The number of female children the judge has.
`progressive.vote`	The proportion of the judge's votes on women's issues which were decided in a pro-feminist direction.
`race`	The judge's race (`1` = white, `2` = African-American, `3` = Hispanic, `4` = Asian-American).
`religion`	The judge's religion (`1` = Unitarian, `2` = Episcopalian, `3` = Baptist, `4` = Catholic, `5` = Jewish, `7` = Presbyterian, `8` = Protestant, `9` = Congregationalist, `10` = Methodist, `11` = Church of Christ, `16` = Baha'i, `17` = Mormon, `21` = Anglican, `24` = Lutheran, `99` = unknown).
`republican`	Takes a value of `1` if the judge was appointed by a Republican president, `0` otherwise. Used as a proxy for the judge's party.
`sons`	The number of male children the judge has.
`woman`	Takes a value of `1` if the judge is a woman, `0` otherwise.
`X`	Indicator for the observation number.
`yearb`	The year the judge was born.

Note: if you are asked to provide an interpretation or an explanation, 1-3 sentences are sufficient. There is no need to write more than that. Short, succinct answer are preferred.

Question 5 (3 points)

Load the `dbj.csv` file. Our outcome in this exercise will be the proportion of pro-feminist decisions, `progressive.vote`. What is the difference in the proportion of pro-feminist decisions between judges who have at least one daughter and those who do not have any? Compute this difference in two ways; (1) using any judge who has children, (2) separately for judges that one, two, or three children. For (2), you should

end with three estimates: one for the judges with one child, one for the judges with two children, and one for the judges with three children (HINT: you will subset the data quite a few times to achieve this).

Answer:

```
library(tidyverse)

dbj <- read_csv("data/dbj.csv") %>%
  mutate(has_children = ifelse(child > 0, 1, 0),
         has_female_children = ifelse(girls>0, 1, 0))

# (1)

dbj %>% filter(has_children == 1) %>%
  group_by(has_female_children) %>%
  summarise(mean_prog = mean(progressive.vote)) %>%
  summarise(diff = mean_prog[has_female_children == 1] -
              mean_prog[has_female_children == 0])
```

```
## # A tibble: 1 x 1
##    diff
##   <dbl>
## 1 0.110
```

```
# (2)

diff_list <- dbj %>%
  filter(child %in% c(1,2,3)) %>%     ## Subset to people w/ 1,2,3 children
  group_by(has_female_children, child) %>%
  summarise(mean_prog = mean(progressive.vote)) %>%  ## Average
  ungroup() %>%
  group_by(child) %>%  ## Calculate differences by no. of children
  summarise(diff = mean_prog[has_female_children == 1] -
              mean_prog[has_female_children == 0])

diff_list
```

```
## # A tibble: 3 x 2
##   child   diff
##   <dbl>  <dbl>
## 1     1 0.241
## 2     2 0.0848
## 3     3 0.189
```

Problem Set 2

Question 5 (3 points)

Creating descriptive statistics is a key part of any research project. We will work with data compiled by Raj Chetty and coauthors. The paper is here: https://www.nber.org/papers/w23618. Alternatively, you can also read a summary report here: https://opportunityinsights.org/wp-content/uploads/2018/03/coll_mrc_summary.pdf). The purpose of this data is measure the role of colleges in facilitating upward (economic) mobility.

1. Using a for loop, calculate the mean and standard deviation of the following variables in the Chetty data: `sat_avg_2013`, `par_median`, `par_top1pc`. The result should be a data frame with 3 columns: `variable`, `mean` and `sd`, and three rows, one for each variable.

** Answer:**

```
library(tidyverse)

df <- read.csv("data/chetty_data.csv")

## Create empty data frame

df_loop <- data.frame(
  variable = character(),
  mean = numeric(),
  sd = numeric()
)

## Create vector of variables

vars <- c("sat_avg_2013", "par_median", "par_top1pc")

## Loop over vars

for (i in vars) {

  ## Calculate mean and sd

  mean_i <- mean(df[[i]], na.rm = T)
  sd_i <- sd(df[[i]], na.rm = T)

  ## Add to df_loop

  df_loop <- df_loop %>%
    add_row(variable = i,
            mean = mean_i,
            sd = sd_i)

}

df_loop
```

```
##       variable         mean           sd
## 1 sat_avg_2013 1.072899e+03 1.389658e+02
## 2   par_median 9.314596e+04 2.908532e+04
## 3   par_top1pc 2.536281e-02 3.624760e-02
```

2. In R, an alternative to for loops are the `apply` family of functions. Repeat subquestion 1 using the

`apply` function of your choice (eg `apply`, `sapply`, `lapply`). The output should be the same as for subquestion 1.

Answer:

```
## Create vector of variables

vars <- c("sat_avg_2013", "par_median", "par_top1pc")

## Loop over vars

list_loop <- lapply(vars, function(x) {

  ## Calculate mean and sd

  mean_i <- mean(df[[x]], na.rm = T)
  sd_i <- sd(df[[x]], na.rm = T)

  ## Add to df_loop

  c(x, mean_i, sd_i)

})

## Note that this creates a list
## We can then convert this to a data frame

df_loop <- list_loop %>%
  reduce(rbind)

colnames(df_loop) <- c("variable", "mean", "sd")
df_loop
```

```
##     variable       mean                 sd
## out "sat_avg_2013" "1072.89914833627"   "138.965755836588"
## elt "par_median"   "93145.9582198002"   "29085.3232984685"
## elt "par_top1pc"   "0.0253628121260873" "0.0362476032133248"
```

```
## This "reduces" the list to a data frame
```

The standard deviation measures dispersion of the data. However, we can also create another measure of the dispersion of the sample mean, called the **standard error of the mean**. It is defined as follows: $s_{\bar{x}} = \frac{s_x}{\sqrt{N}}$, where s_x is the standard deviation of a variable x, and N is the number of observations.

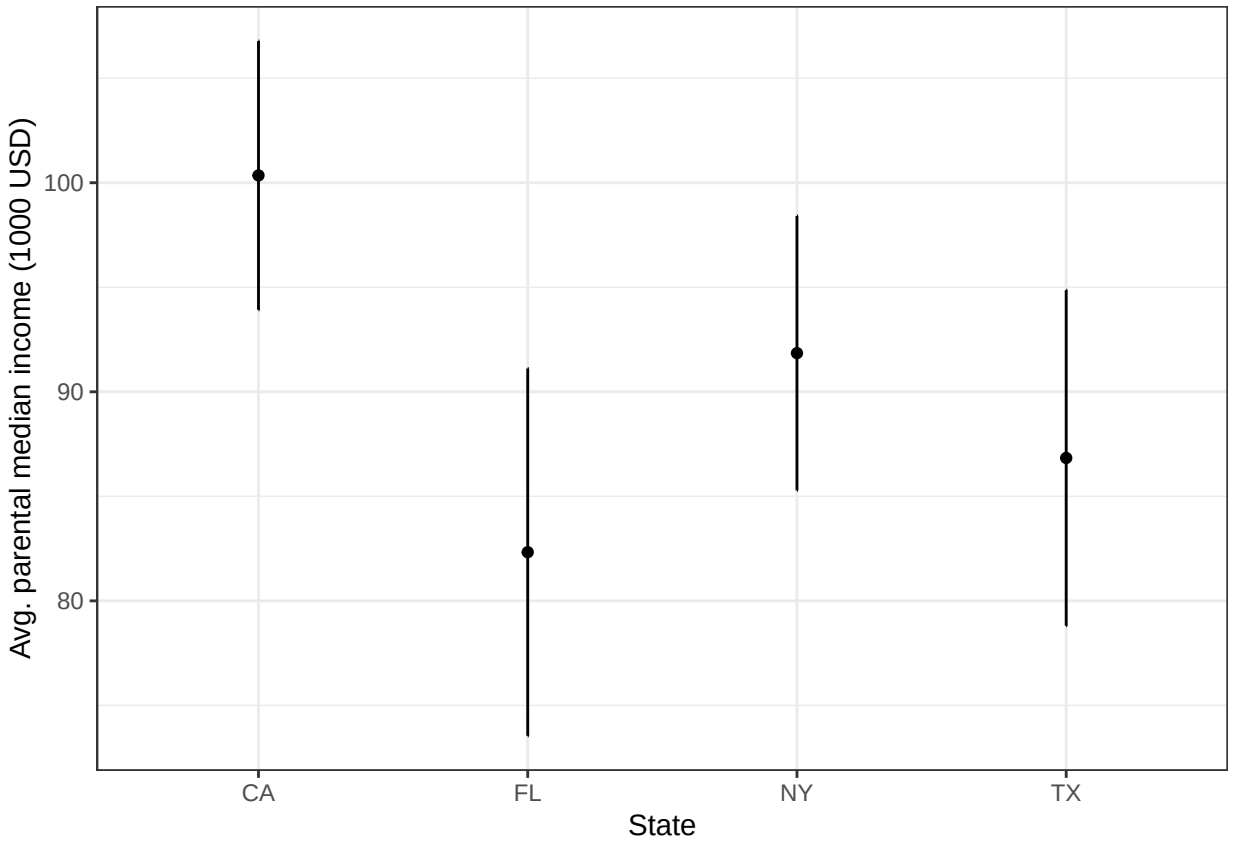
3. Write a short function that calculates the SE of the mean, and takes a variable `x` as the argument. Then, calculate the mean and the SE of the mean for `par_median` separately for CA, NY, TX and FL. Finally, create a plot that has the name of each state on the x-axis. It should then show the mean of `par_median` as a point on the y-axis, and also show error bars that visualize the following interval: $[\bar{x} - 1.96 s_{\bar{x}},\ \bar{x} + 1.96 s_{\bar{x}}]$

- I recommend doing this in `ggplot` using the functions `geom_point` and `geom_errorbar`

Answer:

```
## Function to calculate SE of the mean

my_se_mean <- function(x) {
```

```r
  ## Remove missings from x

  x <- x[!is.na(x)]

  ## SE of the mean

  sd(x) / sqrt(length(x))


}

## Calculate mean and SE of mean by state state

df_plot <- df %>%
  filter(state %in% c("CA", "FL", "TX", "NY")) %>%
  group_by(state) %>%
  summarise(x_bar = mean(par_median),
            x_bar_se = my_se_mean(par_median)) %>%
  ungroup()

## Divide by 1000s for better readability

df_plot <- df_plot %>%
  mutate(across(all_of(c("x_bar", "x_bar_se")), ~./1000))

## Plot this

p1 <- ggplot(df_plot, aes(state, x_bar)) +
  geom_errorbar(aes(ymin = x_bar - 1.96*x_bar_se,
                    ymax = x_bar + 1.96*x_bar_se),
                width = 0) +
  geom_point() +
  theme_bw() +
  ylab("Avg. parental median income (1000 USD)") +
  xlab("State")

p1
```


Avg. parental median income (1000 USD)
100
90
80
CA
FL
NY
TX
State

Problem Set 3

Question 5 (3 points)

5.1

In lecture, we said the mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the best "guess" of x_i in terms of minimizing the mean squared error. In statistics, it is common to define some measure of how "good" a guess or an estimate is, and then to show that there is some specific estimate that is "best" in terms of this measure. Recall that the mean squared error is defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x})^2$$

where $\hat{x}$ is our guess of x_i. [1]

- Show that the mean $\bar{x}$ is the best guess in terms of minimizing the mean squared error given above.
- To show this, you will use calculus. Recall that we can use derivatives to find mimima / maxima of a function. Commonly, we proceed as follows: (1) we take the derivate of a function with respect to the variable we want to optimize over, (2) we set the derivative equal to zero, and (3) we solve for the variable we want to optimize over. In this case, we want to optimize over $\hat{x}$, so we will take the derivative of the mean squared error with respect to $\hat{x}$, set it equal to zero, and solve for $\hat{x}$. The goal is then to show that $\hat{x} = \bar{x}$ is the solution to this problem.
- **Note 1:** Usually, we would also have to assess whether we found a minimum or maximum by looking at the second derivative. However, you can just assume that you found a minimum here – looking at the second derivative is not necessary.
- **Note 2:** You can use the fact that $\sum_{i=1}^{n} a = na$, since a is just a constant.

Answer:

The MSE is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x})^2$$

Note that we can also write this as follows:[2]

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\hat{x} + \hat{x}^2)$$

Next, we take the derivate of this wrt $\hat{x}$:

$$\frac{\partial MSE}{\partial \hat{x}} = \frac{1}{n}\sum_{i=1}^{n}(-2x_i + 2\hat{x})$$

Since we take the derivative wrt $\hat{x}$, we can treat all other terms as constants, which is why the x_i^2 term disappears. Next, we set this equal to zero and solve for $\hat{x}$:

$$\frac{1}{n}\sum_{i=1}^{n}(-2x_i + 2\hat{x}) = 0$$

[1] Note that $\hat{x}$ is usually some function of the data, i.e. $\hat{x} = f(x_1, x_2, ..., x_n)$. Quantities like the mean, median or mode are all functions of the data.

[2] We can also use the chain rule instead, which means we do not need this step.

$$\frac{1}{n}\sum_{i=1}^{n} -2x_i + \frac{1}{n}\sum_{i=1}^{n} 2\hat{x} = 0$$

The second sum is just the sum of constants, i.e. $\sum_{i=1}^{n} 2\hat{x} = 2n\hat{x}$.

$$\frac{1}{n}\sum_{i=1}^{n} -2x_i + \frac{1}{n}2n\hat{x} = \frac{-2}{n}\sum_{i=1}^{n} x_i + 2\hat{x} = 0$$

Note that the first term is negative, so we can just move it to the other side of the equation and divide by 2.

$$2\hat{x} = \frac{2}{n}\sum_{i=1}^{n} x_i$$

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Note that the right-hand side is just the definition of the mean, i.e. we have shown that $\hat{x} = \bar{x}$. As stated above, we would now have to look at the second derivative to see if we found a minimum or maximum. However, you can just assume that we found a minimum here.

5.2

In linear regressions, goodness of fit is usually measured using the R^2 statistic. This statistic is defined as follows:

$$R^2 = 1 - \frac{SSR}{SST}$$

where the sum of squared residuals $SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and the total sum of squares $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$. Here, $\hat{y}_i$ are the predicted values of y_i from the regression, and y_i are the observed values of Y.

In R, write a function that calculates R^2 for a given regression, based on the definition given above. The function should have two arguments: (1) the observed values y_i and (2) an `lm` object that contains the regression results of a regression of Y on X. The function should return the R^2 statistic.

```r
x <- c(2, 5, 3, 6, 7, 8, 3, 6, 7, 9)
y <- c(9, 6, 8, 4, 7, 5, 4, 3, 3, 2)

mod <- lm(y ~ x)

get_rsq  <- function(y, lm_object) {

  # Your code here

}
```

Answer:

```r
get_rsq  <- function(y, lm_object) {

  # Get predicted values

  y_hat <- predict(lm_object)
```

```r
  # Get SSR

  SSR <- sum((y - y_hat)^2)

  # Get SST

  SST <- sum((y - mean(y))^2)

  # Get R^2

  R_sq <- 1 - SSR / SST

  # Return R^2

  return(R_sq)

}

get_rsq(y, mod)
```

```
## [1] 0.4219101
```

5.3

If R^2 is closer to 1, this is usually considered a better fit.

1. First, provide an intuitive explanation why R^2 is a measure of goodness of fit. For this, it is helpful to consider cases where SSR is small relative to SST, and cases where SSR is large relative to SST.
2. Second, please state whether R^2 tells us something about the direction of the relationship between X and Y.
3. Third, assume we are interested in explaining whether citizens vote for Democrats or Republicans. Many different variables are correlated with voting behavior, such as age, income, gender, race, etc. For many of these variables, we can find a "statistically significant"[3] relationship with voting behavior. However, the R^2 of regressing voting behavior on these variables is usually quite low (e.g. < 0.2), which indicates that a given X only explains a small part of the variation in Y. However, researchers in social science generally do not consider this a problem. Please explain why not.

Answer:

1. SST measures the overall deviation of Y from its sample mean. SSR measures the deviation of the predicted values $\hat{y}_i$ from the actual values. If SSR is small relative to SST, then the difference between the predicted and observed values is small, so our model does well. This will then lead to larger values of R^2. If SSR is large relative to SST, then the difference between the predicted and observed values is large, so our model does not do well. This will then lead to smaller values of R^2.
2. Note that the definition of R^2 does not contain X. Therefore, R^2 does not tell us anything about the direction of the relationship between X and Y. It merely tells us whether X is a good predictor of Y.
3. Taken from Gailmard, page 51: "the object of our analysis [...] is typically not to give a complete account of the behavior of Y in a specific sample but rather to identify a set of factors that systematically affect Y and to give reasons behind these effects."

[3] We will cover what this means in more detail later, but I am sure you heard this term before.

Problem Set 4

Question 6 (3 points)

6.1

Assume we have a continuous random variable with the following probability density function (PDF):

$$f(x) = \begin{cases} ax & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Here, a is a constant (i.e. just a number). One requirement for this to be a proper probability density function is that the integral over all possible values of X is equal to 1. In other words, we need to have:

$$\int_0^2 f(x)dx = 1$$

Note that the integral has to be over all possible values of x, which in this case is $0 \leq x \leq 2$.

- First, identify which value the constant a has to have to make this a proper probability density function.
- Then, calculate the expected value of X.

Answer: We can first calculate the value of the integral:

$$\begin{aligned} \int_0^2 f(x)dx &= \int_0^2 axdx \\ &= \left[\frac{ax^2}{2}\right]_0^2 \\ &= 2a \end{aligned}$$

We know that the integral above has to be equal to 1, so we can solve for a. This simply means that $1 = 2a$, which implies that $a = 1/2$.

Now, we can calculate the expected value of X, which does not depend on a:

$$\begin{aligned} E[X] &= \int_0^2 xf(x)dx \\ &= \int_0^2 x\frac{1}{2}xdx \\ &= \left[\frac{1}{2}\frac{x^3}{3}\right]_0^2 \\ &= \frac{4}{3} \end{aligned}$$

6.2

Assume you have the following continuous joint PDF for the variables X and Y

$$f(x,y) = \begin{cases} \frac{3}{2}y^2 & \text{if } 0 \leq x \leq 2 \text{ and } 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- What are the marginal probability density functions of X and Y?
- Given what you found, are the two variables independent?

Hint: When you integrate over one of the two variables in a joint distribution, then the limits of integration should be the range of values that the variable can take on. For example, if you want to integrate over X, then the limits of integration should be $0 \leq x \leq 2$.

Answer:

To get the marginal distribution of X, we need to integrate over all values y. This gives us:

$$\begin{aligned} f_X(x) &= \int_0^1 \frac{3}{2} y^2 dy \\ &= \left[\frac{1}{2} y^3\right]_0^1 \\ &= \frac{1}{2} \end{aligned}$$

We can do the same for Y:

$$\begin{aligned} f_Y(y) &= \int_0^2 \frac{3}{2} y^2 dx \\ &= \left[\frac{3}{2} x y^2\right]_0^2 \\ &= 3y^2 \end{aligned}$$

Note the definition of independence requires that the joint distribution is equal to the product of the marginal distributions. In this case, we have:

$$f(x, y) = \frac{3}{2} y^2 = f_X(x) f_Y(y) = \frac{3}{2} y^2$$

Therefore, we can conclude that X and Y are independent.

6.3

Now, let's look at the following joint distribution. Note that c is just a constant.

$$f(x, y) = \begin{cases} c(x + y^2) & \text{if } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- Find the conditional distribution of X given Y, i.e. $f_{X|Y}(x|y)$.
- Then, calculate the following quantity: $Pr[X < \frac{1}{2} | Y = \frac{1}{2}]$.

Note: Unlike in 6.1, it is not necessary to find the value of c to answer this question. You can just treat c as a constant throughout the calculations. If you do everything correctly, you will see that c eventually disappears from the final expression.

Hint: To calculate the second quantity, you can just plug $Y = 1/2$ into the conditional distribution you found, and use integration to find the desired quantity.

Answer: For this, we first need to find the marginal distribution of Y.

$$f_Y(y) = \int_0^1 c(x+y^2)dx = \left[c\left(\frac{x^2}{2} + xy^2\right)\right]_0^1 = \frac{c}{2} + cy^2$$

Now to find the conditional distribution of X given Y, we need to divide the joint distribution by the marginal distribution of Y:

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f(x,y)}{f_Y(y)} \\ &= \frac{c(x+y^2)}{\frac{c}{2} + cy^2} \\ &= \frac{2(x+y^2)}{1+2y^2} \end{aligned}$$

Next, we want this quantity: $Pr[X < \frac{1}{2}|Y = \frac{1}{2}]$. We can first plug in $Y = 1/2$ into the conditional distribution we found, which gives us:

$$\begin{aligned} f_{X|Y}(x|\frac{1}{2}) &= \frac{2(x+\frac{1}{4})}{1+\frac{1}{2}} \\ &= \frac{4(x+\frac{1}{4})}{3} \end{aligned}$$

We now have the distribution we need. The probability we want is then:

$$\begin{aligned} Pr[X < \frac{1}{2}|Y = \frac{1}{2}] &= \int_0^{\frac{1}{2}} \frac{4(x+\frac{1}{4})}{3}dx \\ &= \left[\frac{4}{3}\left(\frac{x^2}{2} + \frac{x}{4}\right)\right]_0^{\frac{1}{2}} \\ &= \frac{1}{3} \end{aligned}$$

Note that this is just the same as the "usual" way of calculating probabilities for continuous distributions, i.e. we evaluate the integral between the minimum of X and the value we are interested in ($\frac{1}{2}$). The only difference is that we use the conditional distribution instead of the marginal distribution.

Finally, note that we could also have calculated a more general version of this where we plug in $Y = y$ instead of $Y = 1/2$ into the conditional distribution. This would give us:

$$\begin{aligned} Pr[X < \frac{1}{2}|Y = y] &= \int_0^{\frac{1}{2}} \frac{4(x+y^2)}{3}dx \\ &= \left[\frac{4}{3}\left(\frac{x^2}{2} + xy^2\right)\right]_0^{\frac{1}{2}} \\ &= \frac{2}{3}\left(\frac{1}{4} + y^2\right) \end{aligned}$$

The nice thing about this expression is that we can can now plug in any value of y we want. Also, we see that the probability that X is smaller than $\frac{1}{2}$ is increasing in y.

Problem Set 5

Question 5 (3 points)

5.1 (1 point)

Assume that we have a simple linear regression of the form $E(Y|X) = \alpha + \beta X$. The linear regression gives us an estimate of the expected value of Y given X.

Note that the coefficients α and β are chosen to minimize the following expression:

$$E[(Y - \alpha - \beta X)^2]$$

In other words, the coefficients are chosen to minimize the expectation of the squared difference between the expectation of Y given X and the actual value of Y. We can use the last expression to derive the following expression for β and α:

$$\beta = \frac{Cov(X, Y)}{Var(X)}$$
$$\alpha = E(Y) - \beta E(X)$$

Show that the expressions of β and α given above minimize the expression $E[(Y - \alpha - \beta X)^2]$. To do this, you can take the derivatives of the expression in (1) with respect to α and β, and set the derivatives to 0. You can then solve for α and β.[4]

Note: For this question, you can assume the following:

$$\frac{\partial E[f(x, y)]}{\partial(x)} = E\left[\frac{\partial f(x, y)}{\partial x}\right]$$

Where $\frac{\partial f(x,y)}{\partial x}$ is the partial derivative of some function $f(x, y)$ with respect to x.

Hint 1: For some constant a and RVs X and Y:

- $E(aX) = aE(X)$
- $E(a) = a$
- $E(Y + X) = E(Y) + E(X)$

Hint 2: $Var(X) = E(X^2) - E(X)^2$

Hint 3: $Cov(X) = E(XY) - E(X)E(Y)$

Answer:

Let's first define that $e = E[(Y - \alpha - \beta X)^2]$. We then take the derivative of e with respect to α and β.

$$\frac{\partial e}{\partial \alpha} = E[-2(Y - \alpha - \beta X)]$$
$$\frac{\partial e}{\partial \beta} = E[-2X(Y - \alpha - \beta X)]$$

We can now set these to 0. Note that we can take the -2 out of the expectation, since $E(aX) = aE(X)$. Let's start with the partial derivative wrt α:

[4] As in a previous problem set, we will ignore the second derivative, i.e. we will assume we found a minimum.

$$
\begin{aligned}
-2E[(Y - \alpha - \beta X)] &= 0 \\
E(Y) - \alpha - \beta E(X) &= 0 \\
E(Y) - \beta E(X) &= \alpha
\end{aligned}
$$

Note that α and β are constants, which is why we can do the last two steps.

Next, let's look at the derivative of e wrt β:

$$
\begin{aligned}
0 &= E[-2X(Y - \alpha - \beta X)] \\
0 &= -2E[X(Y - \alpha - \beta X)] \\
0 &= E(XY) - \alpha E(X) - \beta E(X^2)
\end{aligned}
$$

We also have a definition of α from above, which we can plug into the equation:

$$
\begin{aligned}
0 &= E(XY) - [E(Y) - \beta E(X)]E(X) - \beta E(X^2) \\
0 &= E(XY) - E(X)E(Y) + \beta[E(X)^2 - E(X^2)] \\
\beta[E(X^2) - E(X)^2] &= E(XY) - E(X)E(Y) \\
\beta &= \frac{E(XY) - E(X)E(Y)}{E(X^2) - E(X)^2}
\end{aligned}
$$

Given what is stated in the hints, the last expression is equal to:

$$
\beta = \frac{Cov(X, Y)}{Var(X)}
$$

In addition, we previously showed that:

$$
\alpha = E(Y) - \beta E(X)
$$

5.2 (0.5 points)

Next, recall that the definition of the correlation coefficient is:

$$
\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}
$$

- Mathematically, what is the relationship between the correlation coefficient and the slope of the bivariate linear regression?
- When are the two the same?
- Can you provide a brief intuition / explanation for the condition under which the two are the same?

Answer:

From the definition of the correlation coefficient, we know that:

$$
Cov(Y, Y) = \rho(X, Y)\sqrt{Var(X)Var(Y)}
$$

We can now substitute this into the expression for β:

$$\begin{aligned}\beta &= \frac{Cov(X,Y)}{Var(X)} \\ &= \frac{\rho(X,Y)\sqrt{Var(X)Var(Y)}}{Var(X)} \\ &= \rho(X,Y)\frac{\sqrt{Var(X)Var(Y)}}{Var(X)} \\ &= \rho(X,Y)\sqrt{\frac{Var(Y)}{Var(X)}}\end{aligned}$$

- The two are the same when $\sqrt{\frac{Var(Y)}{Var(X)}} = 1$, which is the case when $Var(Y) = Var(X)$.
- Intuitively, the condition $Var(Y) = Var(X)$ implies that the dispersion of the variables is the same, which can be interpreted as the two variables being on the same scale.

5.3 (0.5 points)

Consider the following PDF:

$$f(x) = \begin{cases} \frac{3}{2}x^2 & \text{if } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

a. What is $E(X)$? Please do not use the integrals here, but rather come up with an intuitive explanation based on the shape of the distribution.

b. What is the expected value of the new random variable Y, which is defined as the absolute value of X, i.e. $Y = |X|$?

Answer:

a. We can see that the PDF is symmetric around 0. This means that that the area under the PDF to the left of 0 is the same as the area under the PDF to the right of 0. As a result, the expected value of X is 0.

b. We can now calculate the expected value of Y. First, note that Y is a function of X, so we can use LOTUS to calculate the expected value of Y. So we can just use the PDF of X to calculate the expected value of Y.

Next, note that $Y = |X|$ is a function that is defined as follows:

$$Y = \begin{cases} -X & \text{if } X < 0 \\ X & \text{if } X \geq 0 \end{cases}$$

Note, since $Y = |X|$, Y takes the same value for both negative and positive values pf X (e.g. for $X = -0.5$ and also for $X = 0.5$, $Y = 0.5$). Since the PDF of X is symetric, we also know that $f(x) = f(-x)$ for some x between 0 and 1. Therefore, expectation Y for the "positive half" of X should be the same as the expectation for the "negative half" of X. We therefore just calculate the expectation for the "positive half" of X and multiply it by 2:

$$
\begin{aligned}
E(Y) &= 2\int_0^1 x \cdot f(x)dx \\
&= 2\int_0^1 x \cdot \frac{3}{2}x^2 dx \\
&= 3\int_0^1 x^3 dx \\
&= 3\left[\frac{x^4}{4}\right]_0^1 \\
&= 3\left[\frac{1}{4} - \frac{0}{4}\right] \\
&= \frac{3}{4}
\end{aligned}
$$

5.4 (0.5 points)

Suppose that a point is chosen at random on a stick of length 1 and that the stick is broken into two pieces at that point. Find the expected value of the length of the longer piece.

Answer:

The breakpoint is a uniform random variable:

$$X \sim U(0,1)$$

If $X \leq 0.5$, the longer piece is $1 - X$. If $X > 0.5$, the longer piece is X. Note that, by the definition of the uniform distribution, the two cases are equally likely. We can define a new random variable Y that tells us the length of the longer piece:

$$
Y = \begin{cases} 1 - X & \text{if } X \leq 0.5 \\ X & \text{if } X > 0.5 \end{cases}
$$

Note that Y is a function of X. We can therefore use LOTUS to calculate the expected value of Y:

$$
\begin{aligned}
&= \int_0^{0.5} (1-x) \cdot 1dx + \int_{0.5}^1 x \cdot 1dx \\
&= \left[x - \frac{x^2}{2}\right]_0^{0.5} + \left[\frac{x^2}{2}\right]_{0.5}^1 \\
&= \left[0.5 - \frac{0.5^2}{2}\right] + \left[\frac{1^2}{2} - \frac{0.5^2}{2}\right] \\
&= 0.5 - \frac{0.5^2}{2} + \frac{1^2}{2} - \frac{0.5^2}{2} = 0.75
\end{aligned}
$$

Note that we use LOTUS since $Y = g(X)$, so we can just use the PDF $f(x) = 1$ even though we don't consider X directly (however, for the case $X > 0.5$, $g(X)$ is simply X).

Alternatively, this can also be deduced without integration as follows:

- If $X > 0.5$, then Y is a uniform RV on the interval $[0.5, 1]$. The expected value of a uniform RV on the interval $[a, b]$ is $\frac{a+b}{2}$. Therefore, $E(Y|X > 0.5) = \frac{0.5+1}{2} = 0.75$.
- If $X \leq 0.5$, then Y is also a uniform RV on the interval $[0.5, 1]$. Since X can be at most 0.5, Y can only range from 0.5 to 1. Therefore, $E(Y|X \leq 0.5) = \frac{0.5+1}{2} = 0.75$.

So for both $X > 0.5$ and $X \leq 0.5$, $E(Y|X) = 0.75$. As a result, the expected value of Y is also 0.75.

5.5 (0.5 points)

Use a simulation to check your answer to question 5.4.

- First, simulate 1,000 draws to verify whether you answer from the previous question is correct.
- Then, write code that simulates $\bar{Y}_n$, which is the mean of Y as a function of the number of draws n from the uniform RV X. Here, $n \in \{10, 20, 30, \ldots, 500\}$. What do you notice about the relation between $\bar{Y}_n$ and n, in particular with respect to $E(Y)$, which you derived in question 5.4?

Answer:

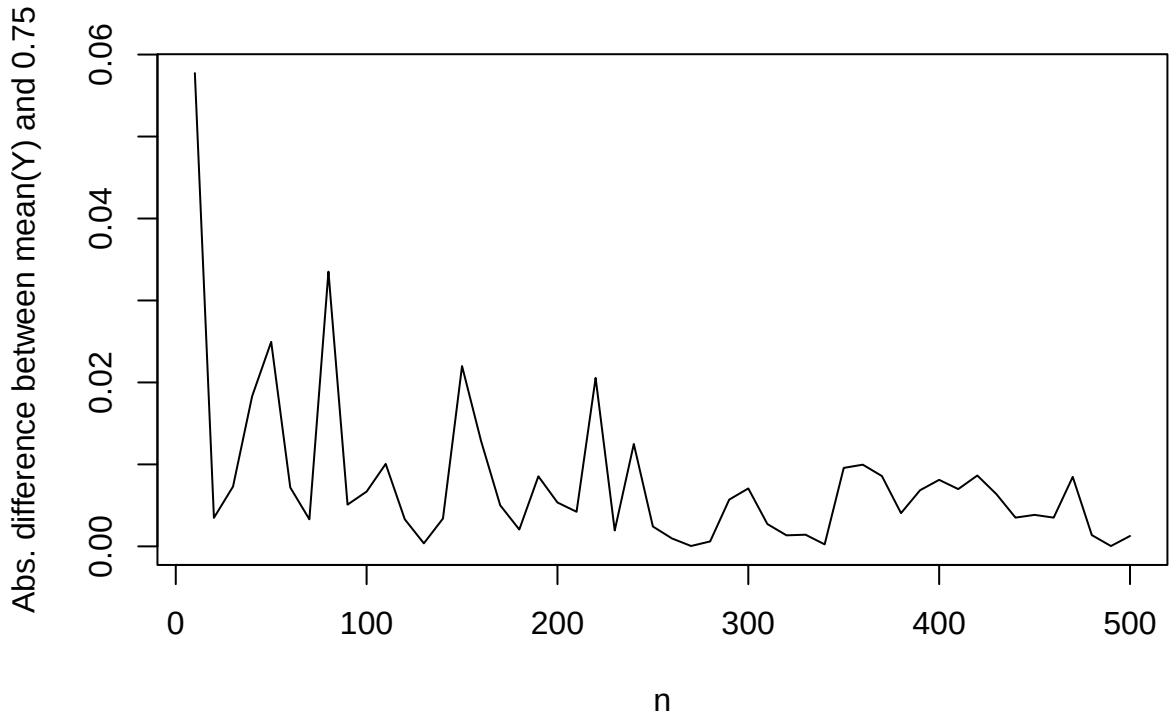
For question 5.4, we can do the following:

```
# Set seed
set.seed(3)

# Draw 1000 values from a uniform distribution on the interval [0,1]
x <- runif(1000, min = 0, max = 1)

# Calculate the length of the longer piece
y <- ifelse(x <= 0.5, 1 - x, x)

# Calculate the mean of the length of the longer piece across all draws
mean(y)
```

```
## [1] 0.7521377
```

```
## Next, do this for n = 10, 20, 30, ..., 500

# Set seed

set.seed(3)

## Function to simulate E(Y_n) given n

simulate <- function(n) {

  # Draw n values from a uniform distribution on the interval [0,1]
  x <- runif(n, min = 0, max = 1)

  # Calculate the length of the longer piece
  y <- ifelse(x <= 0.5, 1 - x, x)

  # Calculate the mean of the length of the longer piece across all draws
  mean(y)

}

## List of n

n <- seq(10, 500, by = 10)

## Simulate E(Y_n) for each n

y_bar <- sapply(n, simulate)
```

```r
## To df

df <- data.frame(n, y_bar)
df$diff = abs(df$y_bar - 0.75)

## Difference between E and 0.75 as a function of n

plot(df$n, df$diff, type = "l", xlab = "n", ylab = "Abs. difference between mean(Y) and 0.75")
```



The first simulation shows that the simulated answer is very close to the analytical answer. The second simulation shows that the mean of Y_n converges to 0.75 as n gets larger.

Probelm Set 6

Question 7 (3 points)

7.1 (1 point)

An important result in statistics is the weak large of large numbers (WLLN), which we will now prove. It states the following:

$$\lim_{n\to\infty} P(|\bar{X}_N - \mu| > \epsilon) = 0$$

where $\bar{X}_N$ is the sample mean of a sample with N iid draws from an RV X with population mean μ. The WLLN say that the probability that the sample mean is more than ϵ away from the population mean goes to zero as the sample size goes to infinity. This is called convergence in probability.

a. As a first step, prove the following, where $\bar{X}_N$ is the sample mean of a sample with N draws from an RV X with population mean μ and finite variance σ^2:

$$\lim_{n\to\infty} E[(\bar{X}_N - \mu)^2] = 0$$

Hint 1: You can use the fact that $E(\bar{X}_N) = \mu$

Hint 2: The proof is short.

b. An important inequality in probability theory is Chebyshev's inequality. It states the following:

$$P(|Y| \geq \epsilon) \leq \frac{E(Y^2)}{\epsilon^2}$$

Here, Y is a random variable and ϵ is a positive number.

Use Chebyshev's inequality and the result from **a.** to prove the weak law of large numbers, which is stated above.

Finally, briefly explain why the variance of the original RV has to be finite for the weak law of large numbers to hold.

Answer:

a. We can use the fact that $Var(\bar{X}) = E[(\bar{X}_N - E(\bar{X}_N))^2] = E[(\bar{X} - \mu)^2]$. We know that $Var(\bar{X}) = \frac{\sigma^2}{N}$. So the limit is equal to $\lim_{n\to\infty} \frac{\sigma^2}{N}$. As N goes to infinity, this limit goes to zero, since σ^2 is a constant.

b. To recap, we want to show the following:

$$\lim_{n\to\infty} P(|\bar{X}_N - \mu| > \epsilon) = 0$$

Aka we want to show that the probability that the sample mean is more than ϵ away from the population mean goes to zero as the sample size goes to infinity.

Now, $|\bar{X}_N - \mu|$ is a random variable, so we can apply the Chebyshev inequality to it. We get:

$$P(|\bar{X}_N - \mu| > \epsilon) \leq \frac{E[(\bar{X}_N - \mu)^2]}{\epsilon^2}$$

Now, assume that N is large. In that case, the RHS is going to converge to zero. This is because the numerator of the RHS is equal to $\frac{\sigma^2}{N}$, which goes to zero as N goes to infinity, and the denominator is a constant.

Since the RHS converges to zero, and the RHS is an upper bound for the LHS, the LHS also has to converge to zero. This proves the WLLN.

Note that the variance of the original RV has to be finite because otherwise the variance of the sample mean would not converge to zero as N goes to infinity. This is because the variance of the sample mean is equal to $\frac{\sigma^2}{N}$, where σ^2 is the variance of the original RV.

7.2 (1 point)

Assume you have two independent uniform RVs distributed as follows:

$$Y \sim \text{Unif}(0, 1)$$
$$X \sim \text{Unif}(1, 2)$$

Assume we draw a sample from each, where the sample size of each sample is N. We are now interested in $\bar{X} - \bar{Y}$.

a. If you want to ensure that, in 95% of the cases, your estimate of $\bar{X} - \bar{Y}$ is within 0.1 of the population difference (i.e. $E(X) - E(Y)$), what is the minimum required sample size N to achieve this?

b. Verify your answer using a simulation in R.

Answer:

a.: The expected value of $\bar{X} - \bar{Y}$ is equal to $E(X) - E(Y) = 1.5 - 0.5 = 1$. The variance of the two sample means is the same:

$$Var(\bar{X}) = Var(\bar{Y}) = \frac{1}{12N}$$

This follows since we know that the variance of the sample mean is the variance of the RV divided by the sample size.

Now, for independent RV, the variance of the sum is equal to the sum of the variances. Therefore, the variance of $\bar{X} - \bar{Y}$ is equal to $\frac{2}{12N}$. We therefore know that the difference between the sample means follows a normal distribution with mean 1 and variance $\frac{2}{12N}$.

For a normal distribution, we know that 95% of the probability mass is within (approx.) 1.96 standard deviations of the mean. As a result, the square root of the variance of $\bar{X} - \bar{Y}$ has to be less than 0.1/1.96. This gives us the following inequality:

$$\sqrt{\frac{2}{12N}} \leq \frac{0.1}{1.96}$$

We can now solve this for N:

$$\begin{aligned}
\sqrt{\frac{2}{12N}} &\leq \frac{0.1}{1.96} \\
\frac{2}{12N} &\leq \frac{0.1^2}{1.96^2} \\
\frac{2}{12} &\leq \frac{0.1^2}{1.96^2} \times N \\
\frac{2}{12} \times \frac{1.96^2}{0.1^2} &\leq N \\
N &\geq 64.02667
\end{aligned}$$

b.

```r
set.seed(123)

## Function to calculate the difference between the sample means of two uniform RVs

get_diff <- function() {
  x <- runif(64, 0, 1)
  y <- runif(64, 1, 2)

  mean(y) - mean(x)
}

## Simulate 10,000 draws from the distribution of the difference between the sample means

out <- replicate(10000, get_diff())

## Calculate the share of draws that are *not* within 0.1 of the population difference

mean(out < 0.9)
```

```
## [1] 0.0247
```

```r
mean(out > 1.1)
```

```
## [1] 0.026
```

```r
## We expected these to be about 2.5% each, which is the case
```

7.3 (1 point)

For a linear regression model with observed values x_i and y_i, the least squares estimator for the slope β is given by:

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of x_i and y_i, respectively. Let's define $S_X^2 = \sum_{i=1}^{N}(x_i - \bar{x})^2$. In addition, the variance of the RV Y is $sigma^2$, and the draws y_i are iid. Note that we can also write the above equation as:

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})y_i}{S_X^2}$$

Find the variance of the least squares estimator $\hat{\beta}$. Note that for the purposes of this problem, we can treat the values x_i as constants, i.e. not as random variables.[5]

Answer:

We can use the fact that $Var(ay_1 + by_2) = a^2Var(y_1) + b^2Var(y_2)$ for any two independent RVs y_1 and y_2 (we assume those to be independent) and constants a and b. Then:

[5] This is a convention that is often used in statistics – we consider β to be a function of the RV Y, but not of the RV X.

$$
\begin{aligned}
Var(\hat{\beta}) &= Var\left(\frac{\sum_{i=1}^{N}(x_i-\bar{x})y_i}{S_X^2}\right) \\
&= \frac{1}{S_X^4}Var\left(\sum_{i=1}^{N}(x_i-\bar{x})y_i\right) \\
&= \frac{1}{S_X^4}\sum_{i=1}^{N}Var\left[(x_i-\bar{x})y_i\right] \\
&= \frac{1}{S_X^4}\sum_{i=1}^{N}(x_i-\bar{x})^2Var(y_i) \\
&= \frac{1}{S_X^4}\sum_{i=1}^{N}(x_i-\bar{x})^2\sigma^2 \\
&= \frac{\sigma^2}{S_X^4}\sum_{i=1}^{N}(x_i-\bar{x})^2 \\
&= \frac{\sigma^2}{S_X^4}S_X^2 \\
&= \frac{\sigma^2}{S_X^2}
\end{aligned}
$$

Problem Set 7

Question 6 (3 point2)

Q6 builds upon the context established in Q5. The framework set up in Q5 is as follows:

In a randomized experiment, researchers want to evaluate an intervention that informs young people about the importance of participation in local politics. The outcome that the researchers are interested in is the amount of hours that the participants spend participating in local politics, which includes activities such as attending city council meetings, working for local political campaigns, and so on. Individuals are randomly assigned to the intervention (the "treatment group") or to a control group where they are exposed to information unrelated to politics. After the experiment, the researchers observe the that the average number of hours spent participating in local politics is 2.3 hours per week for the treatment group, and 2 hours per week for the control group. The variance of the number of hours spent participating in local politics is 8 for both groups. The sample size for both groups is 400.

For the rest of the problem, let the sample mean for the treatment group be $\bar{X}_T$, and the sample mean for the control group be $\bar{X}_C$.

6.1 (1 point)

Assume the same setup as in question 5, with $N_T = N_C = 900$. We now want to calculate the *power* of the test. Power is defined as the probability of rejecting the null hypothesis when the null hypothesis is false.[6] In particular, we will now assume that the true difference between the population means is 0.3 hours per week, i.e. $\mu_T - \mu_C = \Delta\mu = 0.3$. Further, we choose the significance level to be 0.05.

a. Under the null hypothesis in question 4, what are the values for which we reject the null, using the significance level given above? Show this region graphically in a plot of the distribution of the sample mean under the assumption that the null hypothesis holds.

b. Consider scenario outlined above, where $\Delta\mu = 0.3$. What is the distribution of the difference between the sample means under this scenario? Show this distribution graphically. In the same plot, visualize the rejection region you derived in part **a.**.

c. Under the assumption that the true $\Delta\mu = 0.3$, what is the probability of rejecting the null hypothesis? This is the power of the test. Interpret the quantity you calculated in words.

d. The previous steps only gave us the power for one specific value of $\Delta\mu$. In practice, we are interested in the power for a range of values of $\Delta\mu$. Use R to plot the power of our test for values of $\Delta\mu$ between -0.5 and 0.5. Interpret the shape of the curve.

e. The previous question asked you to show how power depends on the true difference in means $\Delta\mu$. Assuming that $\Delta\mu$ is fixed, what are some other ways of increasing the power of the test described in question 4? Name at least two.

f. We say that we make a type II error if we fail to reject the null when it is false. In the context of this question, what is the probability of making a type II error if $\Delta\mu = 0.3$?

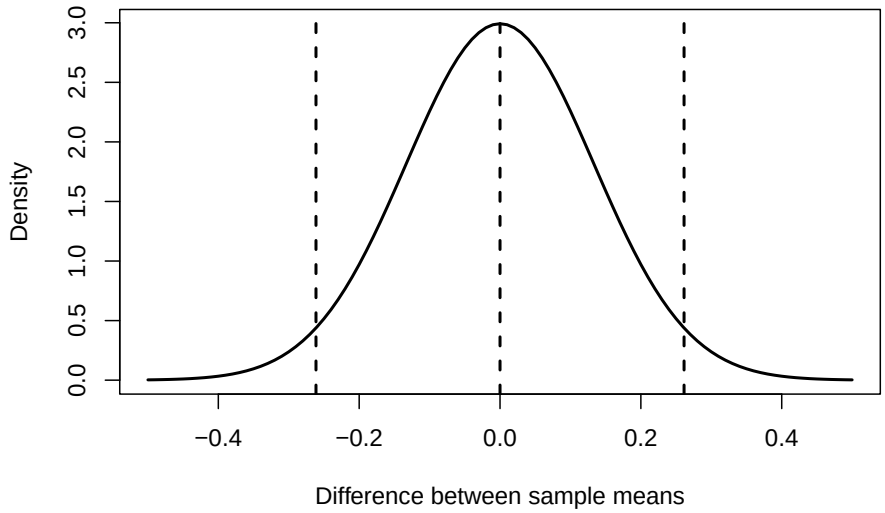
Answers:

a. Under the null hypothesis, the distribution of the difference between the sample means is normal with mean 0 and standard error $\hat{S}_X = \frac{2}{15}$. We reject if either (i) $\Delta\bar{X} < 0 - 1.96\hat{S}_X$ or (ii) $\Delta\bar{X} > 0 + 1.96\hat{S}_X$. Graphically:

```
x <- seq(-0.5, 0.5, 0.01)
y <- dnorm(x, 0, 2 / 15)
plot(x, y,
  type = "l", lwd = 2,
```

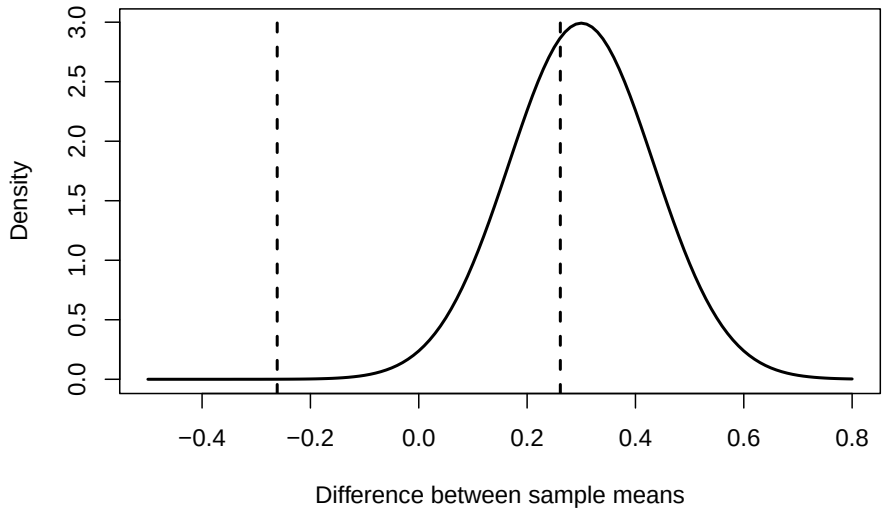
[6] Making statements like "the null hypothesis is false" requires us to make some assumptions about the true population parameters.

```
  xlab = "Difference between sample means",
  ylab = "Density"
)
abline(v = 0, lwd = 2, lty = 2)
abline(v = -1.96 * 2 / 15, lwd = 2, lty = 2)
abline(v = 1.96 * 2 / 15, lwd = 2, lty = 2)
```



b. Under the alternative hypothesis, the distribution of the difference between the sample means is normal with mean 0.3 and standard error $\hat{S}_X = \frac{2}{15}$. Graphically:

```
x <- seq(-0.5, 0.8, 0.01)
y <- dnorm(x, 0.3, 2 / 15)

plot(x, y,
  type = "l", lwd = 2,
  xlab = "Difference between sample means",
  ylab = "Density"
)

abline(v = -1.96 * 2 / 15, lwd = 2, lty = 2)
abline(v = 1.96 * 2 / 15, lwd = 2, lty = 2)
```



c. We already said that we reject the null if the observed difference in means is larger than $1.96 \cdot \frac{2}{15}$ or smaller than $-1.96 \cdot \frac{2}{15}$. Now, under the alternative hypothesis, $\Delta\bar{X} \sim N(0.3, \frac{2}{15})$. We can therefore use the CDF of $\Delta\bar{X}$ under the alternative hypothesis to calculate the probability of rejecting the null hypothesis:

$$P(\Delta\bar{X} \geq 1.96 \cdot \frac{2}{15}) = 1 - F_A(1.96 \cdot \frac{2}{15})$$

To be explicit, we can write F_A to denote the CDF of $\Delta\bar{X}$ under the alternative hypothesis. We can use the normal CDF function in R to calculate this probability:

```
power <- 1 - pnorm(1.96 * 2 / 15, 0.3, 2 / 15)
power
```

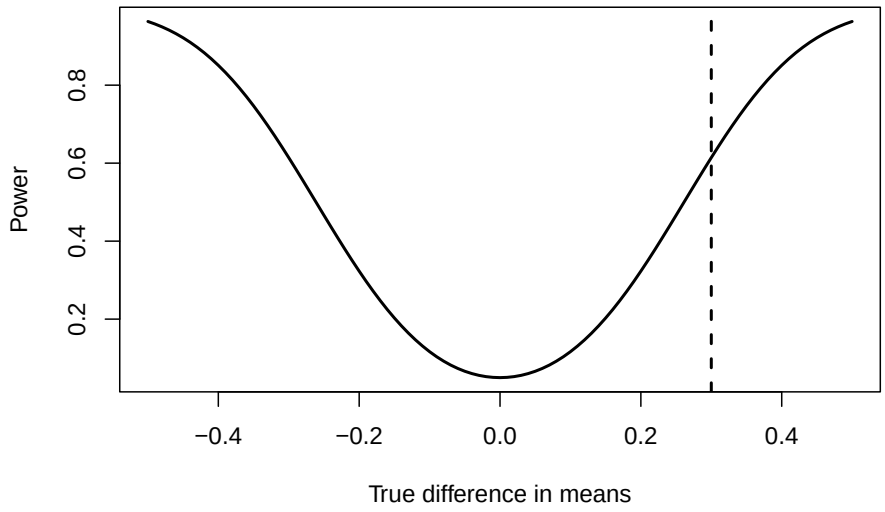
```
## [1] 0.6140919
```

If the true difference in means is 0.3, the probability of rejecting the null hypothesis is about 0.61.

Note that we do not consider the case of $\Delta\bar{X} \leq -1.96 \cdot \frac{2}{15}$, since the likelihood of observing this is practically 0 under the alternative hypothesis.

d. We can use the same approach as in **c.** to calculate the power for a range of values of $\Delta\mu$, which is a function of the CDF of $\Delta\bar{X}$ under the alternative hypothesis. We can use the normal CDF function in R to calculate this probability across values of $\Delta\mu$:

```
delta_mu <- seq(-0.5, 0.5, 0.01)
power <- 1 - pnorm(1.96 * 2 / 15, delta_mu, 2 / 15) +
  pnorm(-1.96 * 2 / 15, delta_mu, 2 / 15)

plot(delta_mu, power,
  type = "l", lwd = 2,
  xlab = "True difference in means",
  ylab = "Power"
)
abline(v = 0.3, lwd = 2, lty = 2)
```



Note that this time, we also consider the rejection region that is smaller than $-1.96 \cdot \frac{2}{15}$, since rejection in this region becomes more likely as we consider smaller / more negative values of $\Delta\mu$.

We can see that, for constant N, the power of the test increases as $|\Delta\mu|$ increases. This makes sense intuitively – we are more likely to correctly reject the null as the true difference in means becomes larger. The original assumed difference in population means from the previous questions is marked by the vertical line.

e. First, we can increase the sample size, which decreases the standard errors and therefore makes the rejection region larger. We can also choose a different significance level (a larger one, i.e. we reject the null hypothesis for a larger range of values of $\Delta\bar{X}$), which also makes the rejection region larger.

f. The probability of making a type II error is 1-power, i.e. the probability of not rejecting the null hypothesis when it is false. If $\Delta\mu = 0.3$, the probability of making a type II error is about 0.39.

6.2

We will continue with the example from question 5. We assume the alternative hypothesis is correct, i.e. $\Delta\mu = 0.3$. We further assume a significance level of 0.05 and standard error of the difference in means that is $SE(\bar{X}) = 4/\sqrt{N}$. However, our goal is now to determine the size of each group N for a one-tailed test that has a power of 0.9.

We now denote $\Delta\mu_0 = 0$ as the difference in means under the null hypothesis, and $\Delta\mu_A = 0.3$ as the difference in means under the alternative hypothesis. For a one-tailed test to have a power of 0.9, the following needs to hold:

$$P\left(\frac{\Delta\bar{X} - \Delta\mu_0}{4/\sqrt{N}} > 1.645\right) = 0.9$$

So we are just saying: the probability of the standardized difference in sample means being greater than the critical value of 1.645 needs to be 0.9. For a one-tailed test with the significance level of 0.05, the critical value is 1.645.

Determine the required sample size N for a power of 0.9 by solving the equation above for N.

Hint 1: This calculation should be done assuming that the alternative hypothesis is true, which implies that $E(\Delta\bar{X}) = \Delta\mu_A = 0.3$.

Hint 2: The inverse of the CDF is called the quantile function. In R, this function is called `qnorm`.

Answer:

$$\begin{aligned}
0.9 &= P\left(\frac{\Delta\bar{X} - \Delta\mu_0}{4/\sqrt{N}} > 1.645\right) = P\left(\frac{\Delta\bar{X}}{4/\sqrt{N}} > 1.645 + \frac{\Delta\mu_0}{4/\sqrt{N}}\right) \\
&= P\left(\frac{\Delta\bar{X} - \Delta\mu_A}{4/\sqrt{N}} > 1.645 + \frac{\Delta\mu_0 - \Delta\mu_A}{4/\sqrt{N}}\right) \\
&= 1 - P\left(\frac{\Delta\bar{X} - \Delta\mu_A}{4/\sqrt{N}} \leq 1.645 + \frac{\Delta\mu_0 - \Delta\mu_A}{4/\sqrt{N}}\right) \\
&= 1 - F\left(1.645 + \frac{\Delta\mu_0 - \Delta\mu_A}{4/\sqrt{N}}\right)
\end{aligned}$$

Here, the we make use of the fact that (i) we can substract $\Delta\mu_A/4/\sqrt{N}$ from both sides of the inequality, and (ii) the expression $(\Delta\bar{X} - \Delta\mu_A)/(4/\sqrt{N})$ is distributed as $N(0,1)$ under the alternative hypothesis.

Finally, the expression in the second to last line is one minus the CDF of a normal distribution with mean 0 and standard deviation 1, evaluated at $1.645 + \frac{\Delta\mu_0 - \Delta\mu_A}{4/\sqrt{N}}$.

Based on the previous derivation, we know that

$$0.1 = F\left(1.645 + \frac{\Delta\mu_0 - \Delta\mu_A}{4/\sqrt{N}}\right)$$

This means that:

$$F^{-1}(0.1) = 1.645 + \frac{\Delta\mu_0 - \Delta\mu_A}{4/\sqrt{N}}$$

Where F^{-1} is the inverse CDF (or the quantile function) of the standard normal distribution. We can now plug in all the values we know (i.e. all values other than N) and solve for N:

$$\begin{aligned}
F^{-1}(0.1) &= 1.645 + \frac{0 - 0.3}{4/\sqrt{N}} \\
\frac{0.3\sqrt{N}}{4} &= F^{-1}(0.1) - 1.645 \\
\sqrt{N} &= \frac{40}{3}\left(F^{-1}(0.1) - 1.645\right) \\
N &= \left(\frac{40}{3}\left(F^{-1}(0.1) - 1.645\right)\right)^2
\end{aligned}$$

We can use the quantile function of the standard normal distribution in R to calculate the value of $F^{-1}(0.1)$:

```
qnorm(0.1)
```

```
## [1] -1.281552
```

This is about -1.281. Plugging this into the equation above, we obtain:

$$N = \left(\frac{40}{3}\left(-1.281 - 1.645\right)\right)^2 \approx 1522.04$$

Problem Set 8

Question 5 (3 points)

An alternative way to test hypotheses without having to make assumptions about the distribution of the test statistic is **randomization inference**.

Let Y_i be a binary outcome for unit i, and D_i be a binary treatment assignment for unit i. As before, the potential outcomes are Y_{1i} and Y_{0i}. We now test a sharp null hypothesis of $H_0 : Y_{1i} = Y_{0i}$ for all i.

For our test statistic, we will use the absolute difference in means estimator, which is defined as follows:

$$T = \left| \frac{1}{N_1} \sum_{i=1}^{N} D_i Y_i - \frac{1}{N_0} \sum_{i=1}^{N} Y_i (1 - D_1) Y_i \right|$$

where N is the total number of units, N_1 is the number of units in the treatment group, and N_0 is the number of units in the control group.

Our data looks like this:

Case	Y_i	D_i
1	1	1
2	1	1
3	1	1
4	0	0
5	1	0
6	0	0

In R, we can calculate the test statistic as follows:

```
# data

Y = c(1, 1, 1, 0, 1, 0)
D = c(1, 1, 1, 0, 0, 0)

# number of units

get_T <- function(Y, D) {

  term1 <- mean(Y[D == 1])
  term2 <- mean(Y[D == 0])

  ## For some possible treatment vectors, the mean of Y[D == 1] or Y[D == 0] may be NaN
  ## In those cases, just set the mean to zero
  ## This is fine since the other term will simply be the mean of Y across all units

  term1 <- ifelse(is.nan(term1), 0, term1)
  term2 <- ifelse(is.nan(term2), 0, term2)

  abs(term1 - term2)

}

get_T(Y, D)
```

```
## [1] 0.6666667
```

The observed T is equal to 2/3.

a. Assume that the null hypothesis holds, but the treatment assignment D_i is different from the one observed above, i.e. some cases receive the treatment that currently do not, and some cases do not receive the treatment that currently do. Under a different treatment assignment, what would the observed values of Y_i be?

b. Assume each unit is able to receive the treatment or not. Let D be a vector that contains the treatment assignment for each unit, which may be different from the one we actually observe. In the observed data, $D = (1, 1, 1, 0, 0, 0)$. How many different vectors D are there in total?

c. Assuming that the null holds, is it possible to calculate T for a different treatment vector D with the observed data given above (i.e. a vector different from the one observed above), or does the fundamental problem of causal inference prevent this?

d. In R, create all possible permutations of the vector D. For each, calculate the test statistic T. This gives you the distribution of T if the null hypothesis is true.

e. We can now calculate a p-value, which tells us how "typical" our observed value of T is under the null hypothesis. Let $\tilde{T}_k$ be test statistic that is derived from the k^{th} permutation of D. The p-value is then defined as follows:

$$p = \frac{1}{K} \sum_{k=1}^{K} 1(\tilde{T}_k \geq T)$$

where K is the total number of permutations, and $1(\cdot)$ is the indicator function.

Calculate the p-value under the null. Using the typical threshold of $\alpha = 0.05$, can you reject the null hypothesis?

f. What is the advantage of randomization inference over the CLT-based approach we have used so far, particularly in cases like the one above?

g. Assume a similar scenario with $N = 4$. In this case, is it possible to reject the null at the 95% confidence level? Why or why not?

h. Unlike in the example above, assume that N is large. Even with a modest sample size of 100, the number of possible permutations of D is $> 10^{30}$. This makes it infeasible to calculate the p-value as we did above. How can we still conduct randomization inference in this case?

Answers:

a. If the null hypothesis holds, the observed values of Y_i would be exactly the same they as they are now, since the unit-level treatment effect is zero.

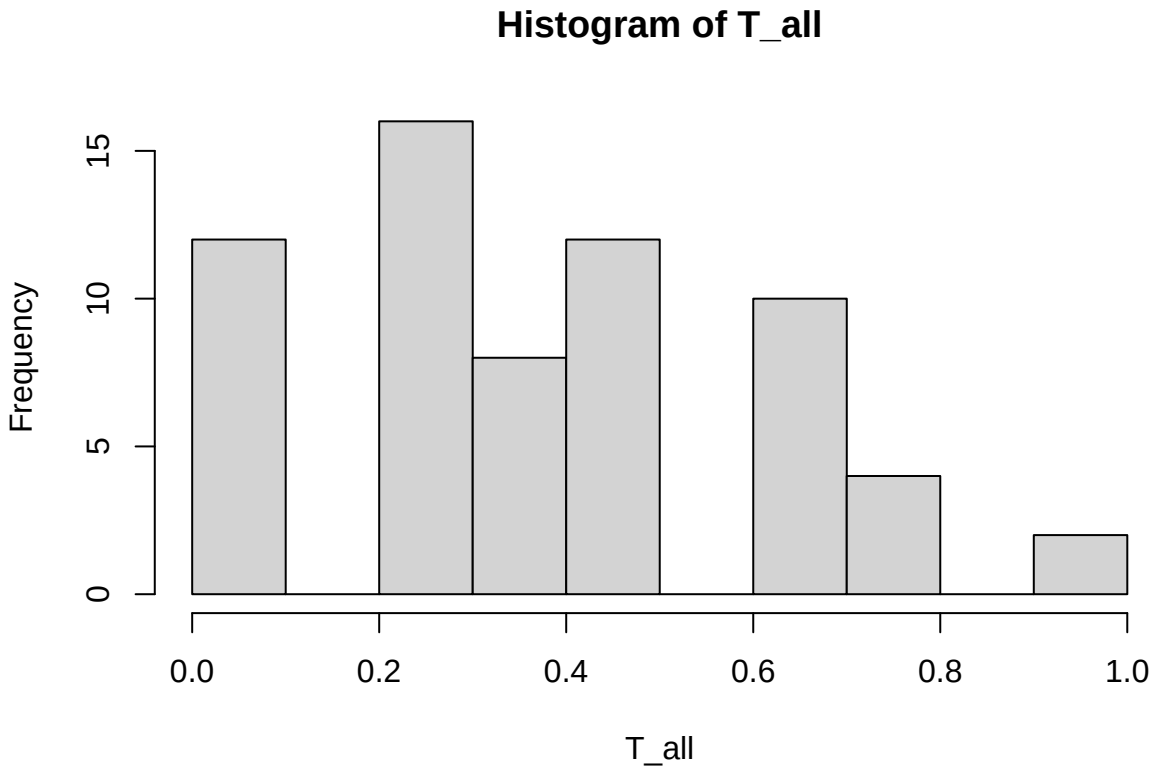
b. There are $2^6 = 64$ different vectors D in total. This is because each unit can either receive the treatment or not, so there are two possibilities for each unit.

c. Under the sharp null, we can calculate T, since the sharp null implies that the observed values of Y_i would be exactly the same they as they are now.

d. We can create all possible permutations of the vector D as follows:

```r
# data

Y = c(1, 1, 1, 0, 1, 0)
D = c(1, 1, 1, 0, 0, 0)

T_obs <- get_T(Y, D)
```

```r
# permutations

library(gtools)

perms <- expand.grid(rep(list(0L:1L), 6L))

## This is a matrix where each row is a permutation of the vector D
## Now, use the function from before to get the distribution of T

T_all <- apply(perms, 1, function(d) get_T(Y, d))

## Distribution

hist(T_all, breaks = 10)
```



e. We can calculate the p-value as follows:

```r
mean(T_all >= T_obs)
```

```
## [1] 0.21875
```

The p-value is 0.25, which means that the observed value of T is not particularly extreme under the null hypothesis. Assuming a threshold of $\alpha = 0.05$, we cannot reject the null hypothesis.

f. The advantage of randomization inference is that it does not require any assumptions about the distribution of the test statistic. In particular, the CLT says that the test statistic follows a normal distribution if the sample size is *large*. This means that the using the normal approximation with small samples (like the one we assess here) may give us incorrect results; randomization inference does not have this problem.

g. If $N = 4$, there are $2^4 = 16$ different vectors D in total. This means that the p-value can be at most $1/16 = 0.0625$. This means that we can never reject the null hypothesis at the 95% confidence level.

h. We can still conduct randomization inference in this case by only assessing a sample of all possible treatment vectors. This is called **permutation testing**. In particular, we can randomly sample K vectors from the set of all possible vectors, and then calculate the test statistic for each. Based on the distribution of the test statistic, we can then calculate a p-value.