Analysis of the randomized plan

---

35 marks

1 Problem

Recall the physical laboratory involving the plane. (See the file background.pdf for more information.)

The equation of the plane can be written as

$$y = \alpha + \beta x + \gamma z$$

which has no error. All of the points lie exactly on the plane and both β and γ are unknown. The planes were arranged so that $\beta = 0.5$ for every run.

Values of x and y were recorded by teams for three different experimental protocols or plans. The third value z was not. In practice, the values of z are never known – there is always some variate that is not measured, perhaps not even thought of, that might be part of relating y to x. These are called **lurking variables** and they will always exist.

The purpose of this question is to investigate and compare the different experimental plans. Of particular interest is whether or not x causes y, that is testing $H_0 : \beta = 0$. And if so, to estimate the value of β defining the causal relationship.

2 Plan

Three experimental plans were considered:

- `"observational"` where six (x, y) pairs were observed in a particular configuration by each team.
- `"randomized"` where three tower-markers were randomly allocated to each of only two different x values, from which the ys were determined. Each team produced two replicates here.
- `"randomizedBlock"` where tower markers were sorted into pairs by height and one marker of each pair were randomly assigned the lower of the two x values and the other to the higher x value. Again, y was determined after allocation for all six markers.

In this question, you will be working only with the data collected using the **randomized** plan. See the file background.pdf for more information.

3 Data

Set up the following:

```r
## Set this up for your own directory
imageDirectory <- "MyAssignmentDirectory/img"  # e.g. in current "./img"
dataDirectory <- "MyAssignmentDirectory/data"  # e.g. in current "./data"
path_concat <- function(path1, ..., sep="/") paste(path1, ..., sep = sep)
```

The full data set is then read in as:

```
labData <- read.csv(file = path_concat(dataDirectory, "labData.csv"))
```

The data can be subsetted according to the three different experimental plans.

a. *(1 mark)* Select that subset of the data corresponding to the randomized plan. Assign it to the variable `randomized`. Show your code.

4 Analysis

b. *(4 marks)* Plot the (x, y) pairs from all of the randomized data

   Use `xlim = c(0, 30), ylim = c(0,40), pch = 19, col = adjustcolor("black", 0.3)` in the call to `plot()`.

   Label the plot meaningfully.

   Fit a straight line model of y on x and add this fitted line to the plot. Save the fit object. Report the value of the slope estimate.

   Show your code.

c. **Learning from repetition.** Each team executed the same plan. Moreover, each team replicated that execution. To gain a better appreciation of the qualities of that plan, we investigate the individual team estimates of β.

   i. *(2 marks)* Separate the data into two subsets, one for each `rep`. Assign the two subsets to the variables `rand1` and `rand2` for replicates 1 and 2. Show your code.

   ii. *(4 marks)* For each replication, fit a separate line for each team's data. For each replication, capture the slope estimates of each team's fit and collect these into a single vector. Call the vector for replication 1's slope estimates `slopes1` and the same for replication 2's `betas2`.

   Show your code.

   ii. *(4 marks)* Plot the $(betas1, betas2)$ pairs from the randomized data

      Use `xlim = c(-1, 1), ylim = c(-1, 1), pch = 19, col = adjustcolor("black", 0.3)` in the call to `plot()`.

      Label the plot meaningfully.

      Show your code.

   iii. *(4 marks)* Test the hypothesis that the team paired slope estimators, $(\tilde{\beta}_1, \tilde{\beta}_2)$, based on replicates 1 and 2, are independently distributed. That is test $H_0 : \tilde{\beta}_1 \perp\!\!\!\perp \tilde{\beta}_2$.

      Use `numericalTest()` with the appropriate choices of discrepancy measure and generation function.

      Show your code.

      Write up your conclusion about the independence.

   iv. *(3 marks)* Draw a meaningfully labelled histogram of the individual slope coefficient estimates for all teams for **replicate 1** only.

      Show your code.

      Use `xlim = c(-1, 1), col = "lightgrey"` in `hist()` and an appropriate `main` title and `xlab`.

      Add a vertical red dashed line at the average of the slope estimates.

      Add a vertical blue dashed line at the true value of β.

      Print the average and standard deviation of the slope estimates.

v. *(3 marks)* Draw a meaningfully labelled histogram of the individual slope coefficient estimates for all teams for **replicate 2** only.

Show your code.

Use `xlim = c(-1, 1), col = "lightgrey"` in `hist()` and an appropriate `main` title and `xlab`.

Add a vertical red dashed line at the average of the slope estimates.

Add a vertical blue dashed line at the true value of β.

Print the average and standard deviation of the slope estimates.

vi. *(3 marks)* For all teams, draw a meaningfully labelled histogram of the average of the two individual slope coefficient estimates (over the two replicates).

Show your code.

Use `xlim = c(-1, 1), col = "lightgrey"` in `hist()` and an appropriate `main` title and `xlab`.

Add a vertical red dashed line at the average of the slope estimates.

Add a vertical blue dashed line at the true value of β.

Print the average and standard deviation of the slope estimates.

5 Conclusion

e. *(3 marks)* What do you conclude about the quality of team slope estimates from the randomized study?

f. *(2 marks)* What do you conclude about the value of having each team average their replicates from the randomized study?

g. *(2 marks)* What effect, if any, has been produced by a lurking variable? Explain.