MINERVA SCHOOLS AT K.G.I.
Assignment 1
Quang Tran
CS112 Spring 2019

TABLE OF CONTENT

Link to code: https://gist.github.com/quangntran/a3fe2bb032c562cac8a02106daf1e21d

Question 1.

You have been told that project duration at approval is generally about 2 years (24 months). In other words, (purportedly) when projects are approved, the difference between the original project completion date and the approval date is (supposedly) approximately 24 months.

(a1) *Is this claim true? Explain.*

The projects with missing values for original project completion date and approval date were removed. The durations (in days) at approval for the remaining projects were calculated. In order for converting days to months, it is assumed that 1 month = 30 days. The mean duration is 21.45897 months, less than the claimed 24 months. However, this difference is probably just by chance and not reflecting the general, true trend of giving a duration of 24 months. To make the conclusion stronger as to whether or not the duration at approval is 24 months, we construct a 95% confidence interval. First, we look at the histogram for the durations at approval of the project.

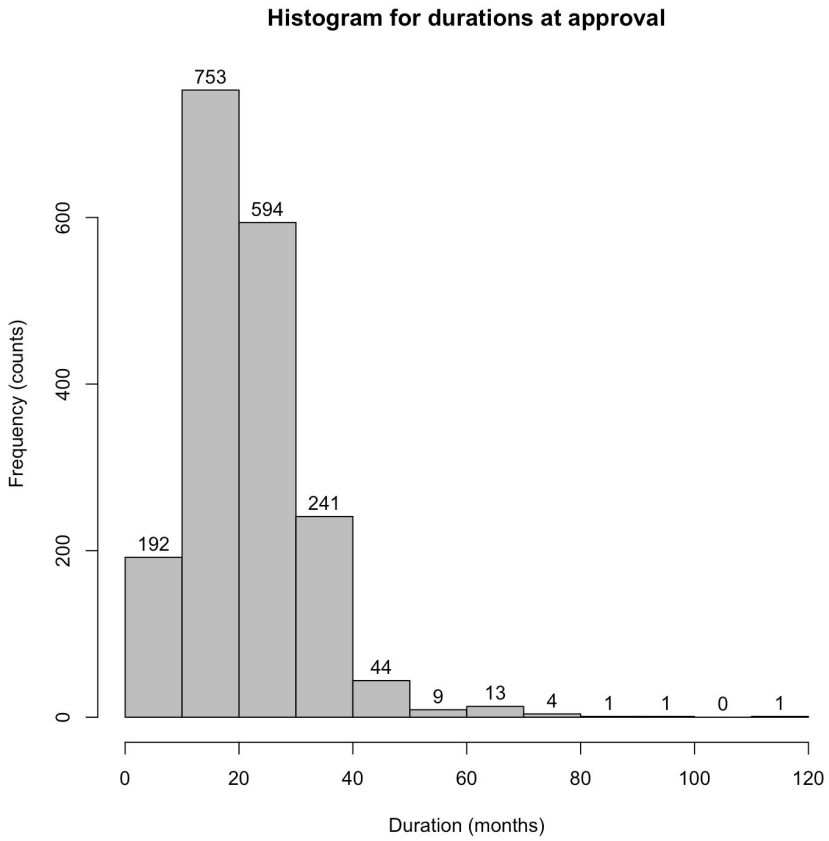

Figure 1. Histogram for durations at approval. The distribution is heavily right-skewed.

As the sample distribution is heavily skewed, we cannot assume normality to the population distribution. Therefore, we will use t-distribution in constructing the confidence interval. The interval is [20.96217; 21.95576]. We see that the interval does not contain the claimed 24 figure, lending no strong evidence that the claim is true.

(a2) *Has project duration at approval changed over time?*

The top 20% early projects are considered to be the early projects, and the top 20% late projects are considered to be the late ones. The durations at approval for those projects are computed and converted to months as above. Below is the density plots for the two groups of projects.

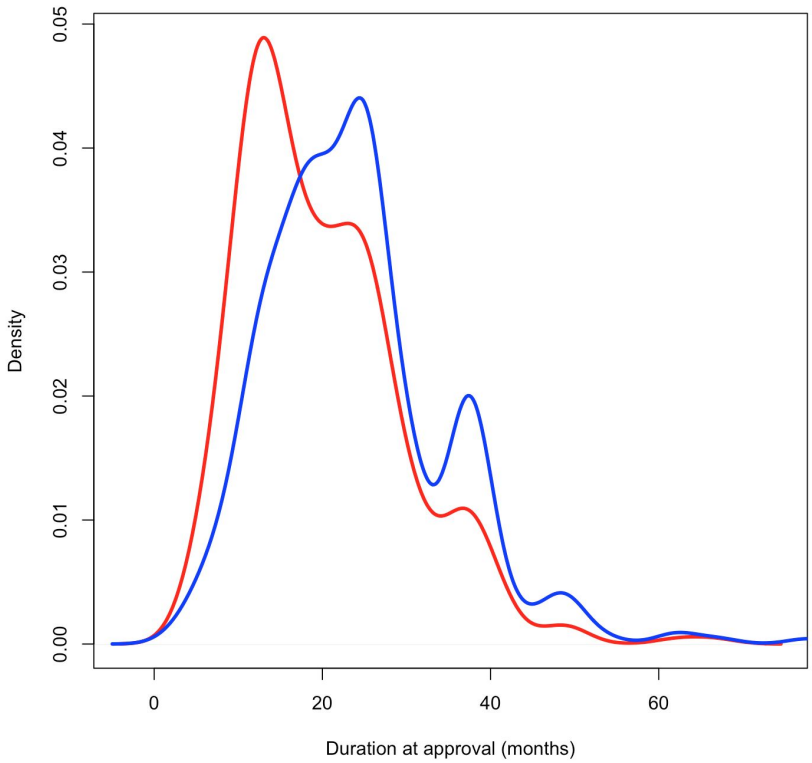

Figure 2. Density plots for early (red) and late (blue) projects.

The mean durations of early and late projects are 19.97 months and 24.16 months, respective. We see that on average, later projects have longer durations at approval. One may argue that the mean is not appropriate for comparing the trends of durations at approval, as it is sensitive to outliers. In fact, the density plots show that there are indeed prominent outliers for both project groups (those with 40+ and 60+ months of duration). Median is “robust” to outliers, so let us have a look at them. The medians for early and late projects are 18.46 months and 23.63 months.

Again, this lends further support that later projects tend to have longer durations at approval. This can be observed in the plot. We see that there is a clear shift to the right from the red line, which corresponds to the early group, to the blue line, which represents the late group.

It is also helpful to look at some measure of spread instead of centrality like mean and median. The interquartile ranges for early and late projects are 12.95 and 12.18, respectively. These are roughly equal figures. As interquartile range measures variability and is robust to outliers (compared to other dispersion measures such as variance and standard deviation), the rough equality means that the way the duration is assigned at approval date is as consistent for the early group as for the late group.

(b) *How does original planned project duration differ from actual duration.*

Again, we reuse the approach in question (a). The mean revised duration is 40.52 months, almost double that of the approval date duration, 21.45 months. We observe the same extent of the difference for the medians: 19.73 months (duration at approval) and 37.3 months (revised duration). These two measures of central tendency of data give support to the claim that the revised durations are generally longer than the original durations.

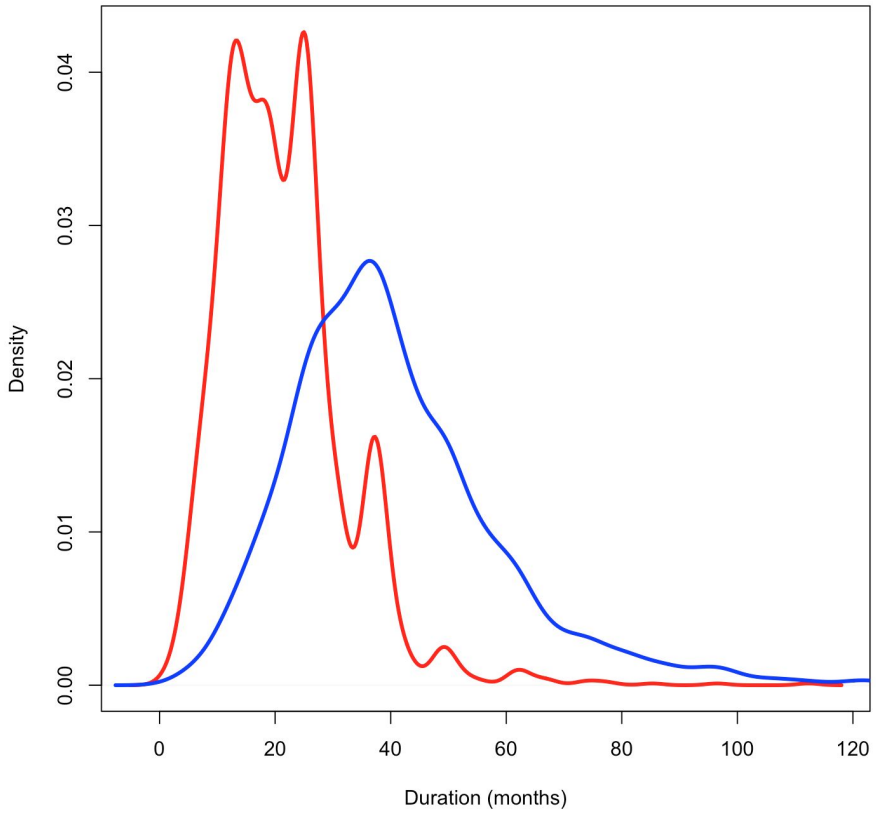

Figure 3. Density plot for durations at approval (red) and revised durations (blue).

The interquartile ranges for the two groups differ by 8.46 (the interquartile range for original durations is 12.8 and that for revised durations is 21.26. This high difference tells us that the revised durations vary to a greater extent than the at-approval durations do. This can be seen in the plot: the red distribution looks “thinner” and the blue line.

Question 2.

What % of projects that have ratings were rated 0? What % were rated 1? What % were rated 2? What % were rated 3? Answer these questions using a table or a figure.

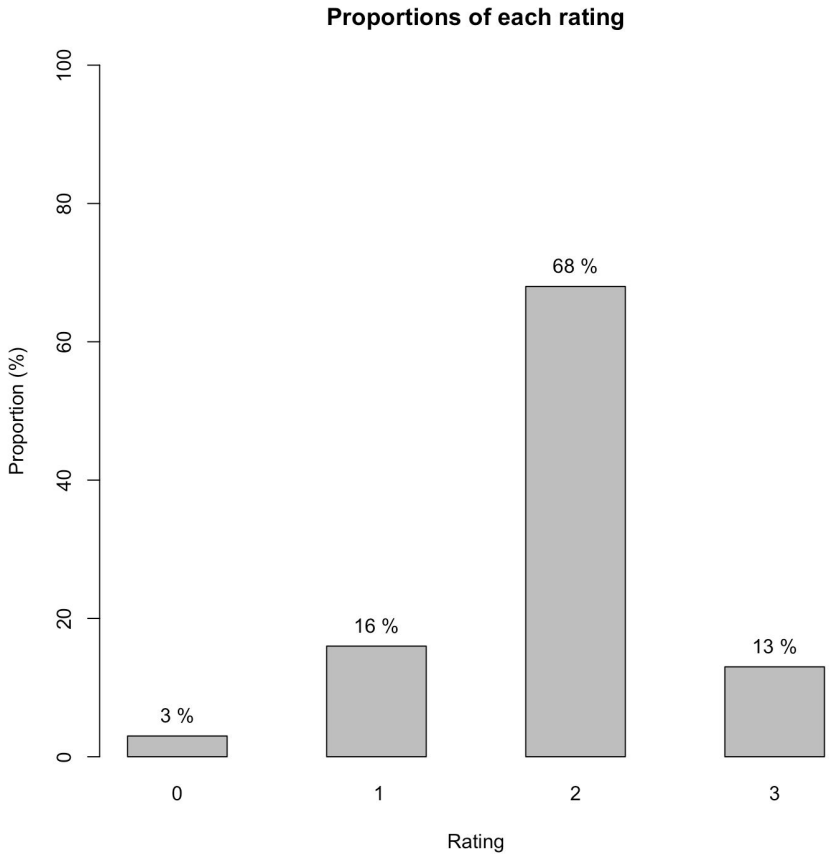

Figure 4. Proportions of each rating. Most projects were rated 2 (68%). Roughly the same percentage of projects received a rate of 1 or 3 (16% and 13%, respectively). Zero rating is the least common one (3%).

Question 3.

Repeat problem 2, but this time exclude all PPTA projects.

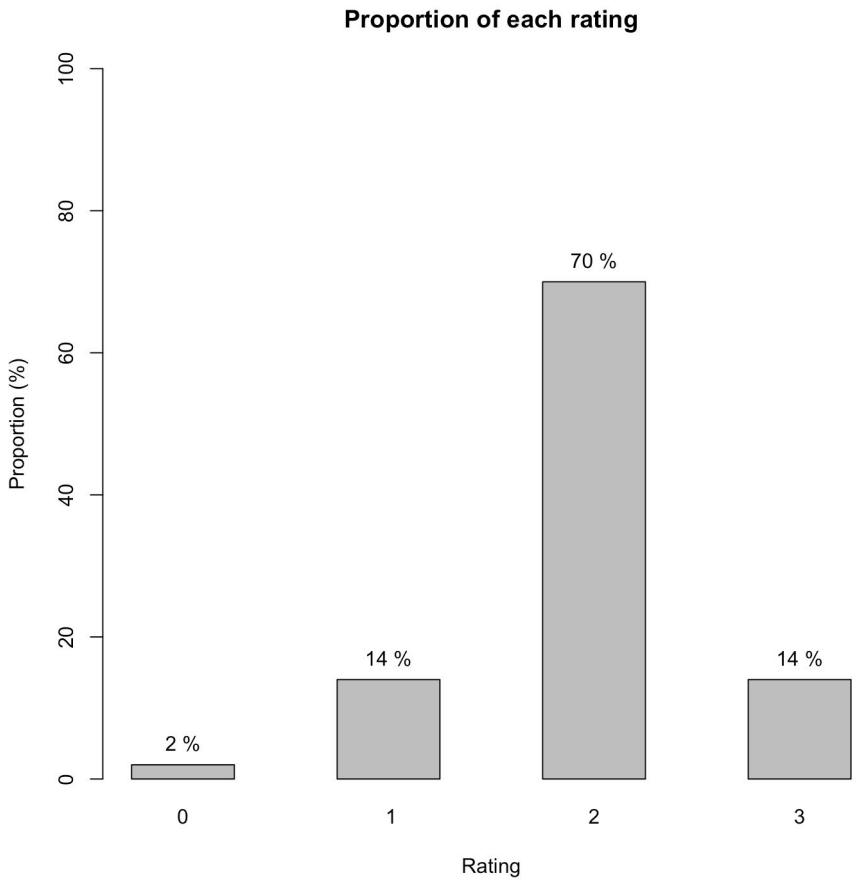

Figure 5. Proportions of each rating when PPTA projects are excluded. The 2 rating remains the most common of the four types.

Question 4.

Identify the top 25% of projects by "Revised.Amount" and the bottom 25% of projects by "RevisedAmount". Compare the ratings of these projects. Can you draw a causal conclusion about the effect of budget size on ratings? Why or why not?

To identify the top 25% of projects by the budget, all projects with budget larger than the 75% quantile is chosen. Similarly, projects with budget smaller than the 25% quantile are the bottom 25%.

The mean rating of the top-budget projects is 1.948, and the mean rating of the bottom ones is 1.945. The difference in means is, therefore, 0.003. The causal effect is very close to 0. It is tempting to say that budgeting has virtually no effect on the success (rating) of a project. However, such a conclusion about the causal relationship is really weak, if not invalid. This is due to several factors the data based on which the causal conclusion is drawn is not randomized (there was no random treatment -- which is the budget -- assignment process). There may be systematic differences between the low- and high-budget group and it could be these differences, not the budget, that are the driving factors for the difference in ratings. For example, the five most common divisions of the top-budget projects are SEPF (23 projects), SEER (20 projects), SDSC (17 projects), SEEN (14 projects), and INRM (13 projects). The same five divisions have very different number of projects associated with them in the bottom projects: while SEPF is also the most common division in the bottom group, there are only 4 SEER projects are SEER, 7 are SDSC, 6 are SEEN, and 6 are INRM. That said, the top 5 common divisions for the bottom group are very different from the top 5 for the top group.

Let us also take a look at the distribution of the prefix of the projects in each group.

Prefix	Cs	RCs	RCsS	CsS	R	RS	S
# Projects	1	1	1	5	41	121	139

Table 1. Prefix distribution for top-budget projects

Prefix	CsS	RCsS	RS	S	Cs	RCs	R
# Projects	0	0	0	0	6	8	202

Table 2. Prefix distribution for low-budget projects

From the two tables we see that while most of the projects have an R prefix for low-budget projects (93.5%), the top two prefixes for top-budget projects are RS and S.

Is the difference in the prefix and the division the cause for the difference in ratings? We never know (at least from this data set), just as we never know if budget is causing the effect.

Question 5.

Imagine your manager asks you to apply Jeremy Howard's drivetrain model to the problem of optimal budget-setting to maximize project success (i.e., "Rating"). In such a situation, what would be the:

(a) *Decision problem or objective?*

How to maximize the rating of any project that comes in the door?
(Here, the objective is the project's rating.)

(b) *Lever or levers?*

Assuming the constraint of the problem is that the budget is the only thing about the project we can manipulate (other dimensions, such as the project's prefix, division, department, etc., are pre-determined), one lever we can pull is the budget set for the project.

(c) *Ideal RCT design?*

Because the treatment, which is the budget set for the project, is a continuous value, for the experiment we can decretize it: creating buckets of budget ($0, $1m, $2m, etc.) The maximum value and the distance between two consecutive buckets should be made as large as possible (ideally speaking, we would assign a project a randomized real, continuous value.)

Suppose we have N combinations over the projects' characteristics (for example, if for any project, there are 3 possible departments, 2 possible divisions, and 2 possible country, then there are 3*2*2=12 possible combinations). We then run N RCTs, each for one combination. In each RCT, we randomly assign (with equal probability) each project in the corresponding combination a treatment bucket (e.g., $1m), and record the ratings. This randomization is to reduce allocation bias. Further rules include:

- The number of samples in each combination is large enough so that the number of projects in each kind of treatment is also sufficiently large.
- The rating system is designed in a way that is blind to the treatments the projects receive to reduce assessment bias (Sibbald & Roland, 1998).
- Any project that does not complete or has not gone far enough to the rating stage is still included in the analysis to prevent attrition bias (Sibbald & Roland, 1998).

(d) *Dependent variable(s) and independent variable(s) in the modeler*

After the RCTs, we have enough data for the modeler. For each group/combination of projects, we model and interpolate the results, then run an optimizer to determine the budget that leads to the best rating for that project group. The independent variable is budget (in dollars), and the dependent variables is rating.

After this step, for every project that comes in the door, we first determine which group it belongs to, based on its characteristics, then choose the optimal budget based on the built optimizer.

(e) *Why would running RCTs and modeling/optimizing over RCT results be preferable to using (observational, non-RCT) "foo" data?*

RCTs involve randomization of treatment assignment along with other measures (listed above) and thus reduce potential sources of bias and confounding variables in causal inference. One key difference we see between the RCTs described above and the "foo" data is that in the RCTs we ensure that the causal effects recorded is attributed to the treatment (the budget) only and not to any other characteristics of group projects because 1) we ensure that, to the best of our current knowledge about the projects' characteristics, the projects in control and treatment groups only differ in the treatments and 2) we randomize the treatment assignment to reduce the effect of confounding variables. It follows that the causal effect in RCTs gives an unbiased estimate of the true causal effect.

References

Sibbald, B., & Roland, M. (1998). Understanding controlled trials: Why are randomised controlled trials important? British Medical Journal, 316(7126), 201. Retrieved December 28, 2015 from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2665449/pdf/9468688.pdf.