Contents

```
require(xtable)
```

```
## Loading required package: xtable
```

```
load("races.RData")
diff.days <- function(time1, time2, ...) {
  as.numeric(difftime(time1, time2, units='days', ...))
}
Total <- function(x) sum(x)
```

1 A comparison in the field

1.1 The experimental setup

In this section, we illustrate the preceding econometric methods by an empirical analysis of the outcomes of a field experiment that we conducted to compare races and tournaments. The context of the experiment was an online programming competition among expert software developers, engineers, and computer scientists. This competition was conducted on the online platform Topcoder.com from March 2 to March 16, 2016.

Since its launch in 2010, Topcoder.com hosts on a weekly basis competitive programming contests for thousands of competitors from all over the world. Typical assigned problems are data science problems (e.g., classification, prediction, natural language processing) that involve some background in machine learning and statistics to be solved. All Topcoder members – about 1M registered users in 2016 – can compete and attain a "rating" that provides a metric of their ability as a competitor. Other than attaining a rating, the competitors having made the top five submissions usually win a monetary prize. The extent of these awards can range considerably depending on the nature and complexity of the problem, generally between $5,000 to $20,000.

In this study, we worked together with researchers from the National Health Institute (NIH) and the Scripps Research Institute (SCRIPPS) to select a challenging problem for an experimental programming competition. The selected problem was based on an algorithm, called BANNER, built by NIH that uses expert labeling to annotate abstracts from PubMed, the most prominent life sciences and biomedical search engine, so disease characteristics can be more easily identified [@leaman2008banner]. The goal of the challenge was to improve upon the current NIH's system by using a combination of expert and non-expert labeling along the lines of @good2014microtask.

The contest was announced to all community members via email. Following a three days registration period, signed up competitors were randomly assigned to separate groups. Each group had to solve the same computational problem, as described above, within a period of 2 weeks.

Groups were then randomly assigned to one of three different competitive settings: a race, a tournament, and a tournament with a minimum quality requirement. In all groups, the first placed competitor was awarded a prize of $1,000, and an additional, consolatory prize of $100 was awarded to the second placed competitor. However, in a race, xxxx. In a tournament, xxxx .And in a tournament with minimum quality requirement, xxxx.

Outcomes were matched with data from each competitor's online web profile on the platform. This typically includes the date in which the member registered, the rating, the number of past competitions, and so on. Additional personal information, was collected via a mandatory initial and a final survey. In the initial survey, registered competitors were asked basic demographics, including a measure of risk aversion. They were also asked to forecast the number of hours they would be able to spend competing in the next few days (the exact question was: "looking ahead xxxx"). In the final survey, they were asked to look back and tell us their best estimate of the time spent working on the problem.

1.2 Data

A total of 299 competitors signed up for the challenge. All were registered members of the platform, with the median time as member of the platform of NA years (min = NA and max = NA years). The distribution of experience in competing was highly skewed, with competitors in the 90th percentile having taken part in 28 more competitions and with a skill rating of 999 higher points than those in the 10th percentile; see Figure 1.

[Figure 1 about here.]

After the two-week submission period, 86 competitors made 1759 submissions overall. The distribution of submissions was rather skewed, with participants in the 90th percentile having made 50 more submissions than those in the 10th percentile.

This result does not seem to correlate well with the competitor's experience or skills, as the Pearsons's correlation coefficient between the count of past competitions or the rating and the count of submissions is positive but generally low; see Table XXX. Thus, differences in submissions appear idiosyncratic and perhaps related to the way to organize the work rather than systematically associated with underlying differences in experience or skills.

```
cor(dat[, c("nsub", "mm.rating", "mm.count")], use='pairwise.complete.obs')
```

```
##                  nsub mm.rating   mm.count
## nsub       1.00000000 0.1147754 0.04634501
## mm.rating 0.11477537 1.0000000 0.33330457
## mm.count  0.04634501 0.3333046 1.00000000
```

The timing of submissions was rather uniform during the submission period with a peak of submissions made in the last of the competition. (explain more)

```
scores$submax <- ave(scores$subid, scores$id, FUN=max)
par(mfrow=c(2, 1), mar=c(4,4,2,2))
plot(subid==1 ~ as.POSIXct(subts), data=scores, type='h', yaxt='n'
    , xlab='', ylab='', main='Dispersion time first submission')
plot(subid==submax ~ as.POSIXct(subts), data=scores, type='h'
    , yaxt='n', xlab='', ylab='', main='Dispersion time last submission')
```

Scores: xxxx

2 Empirical analysis

2.1 Treatment differences

Difference in participation by treatments are show in Table XX.

```
tab <- table(dat$treatment, !is.na(dat$nsub))
fisher.test(tab)
```

```
Fisher's Exact Test for Count Data
```

data: tab p-value = 0.5255 alternative hypothesis: two.sided

```
xtable(addmargins(tab, FUN=Total), digits=0)
```

Margins computed over dimensions in the following order: 1: 2: % latex table generated in R 3.2.3 by xtable 1.8-2 package % Mon Nov 28 16:01:17 2016

[Table 1 about here.]

We find no differences in the room size.

```
tab <- table(dat$size, !is.na(dat$nsub))
fisher.test(tab)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tab
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.5689643 1.6912635
## sample estimates:
## odds ratio
##  0.9846648
```

```
addmargins(tab, FUN=Total)
```

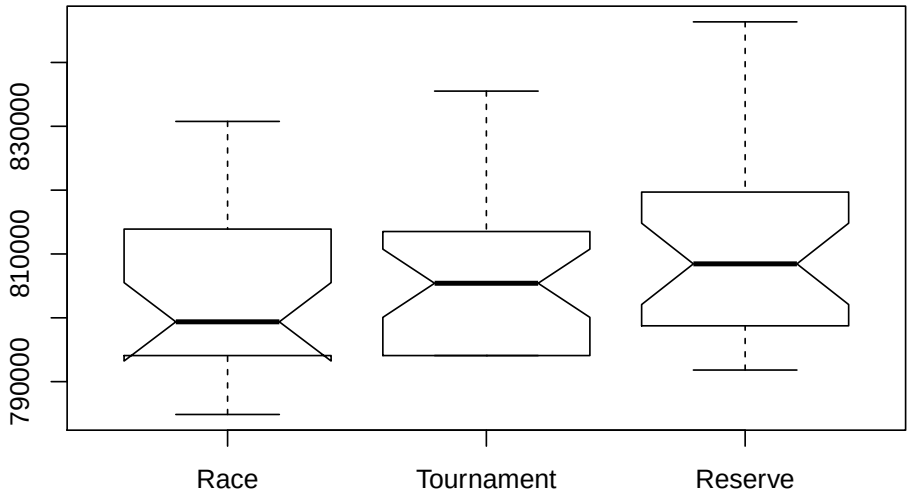
```
## Margins computed over dimensions
## in the following order:
## 1:
## 2:
```

```
##
##         FALSE TRUE Total
##   Large   128   52   180
##   Small    85   34   119
##   Total   213   86   299
```

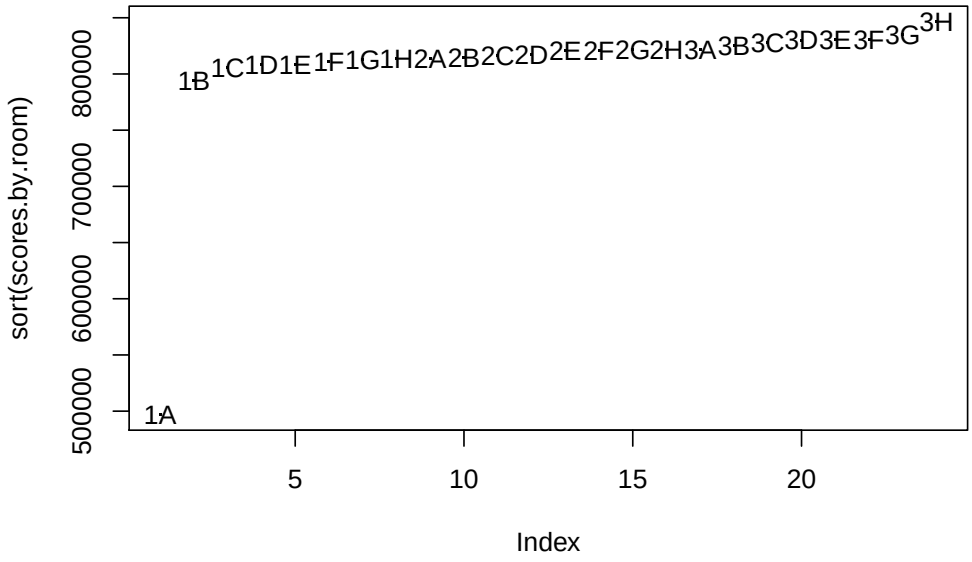
- Ex-post

```
boxplot(dat$lastscore~dat$treatment, outline=FALSE, notch=TRUE)
```

```
## Warning in bxp(structure(list(stats = structure(c(784858.58, 794084.93, :
## some notches went outside hinges ('box'): maybe set notch=FALSE
```




```
scores.by.room <- with(dat, tapply(lastscore, room, max, na.rm=TRUE))
plot(sort(scores.by.room), pch='.')
text(sort(scores.by.room), gsub("Group ", "",names(scores.by.room)))
```



- Timing: early vs late

```
tab <- table(runif(500)<0.5)
# tab <- table(dat$treatment, ifelse(dat$nsub>0, "Submission", "No submission"))
# tab2 <- addmargins(tab, 2, FUN=list(Total=sum))
# print(kable(tab2), type='html')
```

Using a Chi-square test of independence, we find no significant differences in participation rates associated with the assigned treatments (p-value: 0.421); see Table XX.

Further, we model participation rates as a logistic regression. We use a polynomial of third degree for the count of past competitions to account for non-linear effects of experience; and we use an indicator for whether the competitor had a win or not. Also, taking into account differences in ability, participation rates are not significantly different.

2.2 Estimation results

Participation to the competition by treatment is shown in Figure **??**. Participation here is measured by the proportion of registered participants per treatment who made any submission during the eight-day submission period. Recall that competitors may decide to enter into the competition and work on the problem without necessarily submitting. In a tournament, for example, competitors are awarded a prize based on their last submission and may decide to drop out without submitting anything. However, this scenario seems unlikely. In fact, competitors often end up making multiple submissions because by doing so they obtain intermediate feedback via preliminary scoring (see Section XXX for details). In a race, competitors have even stronger incentives to make early submissions as any submission that hits the target first wins.

`Table xxx`

We find that the propensity to make a submission is higher in the Tournament than in the Race and in the Tournament with reserve, but the difference is not statistically significant (a Fisher's exact test gives a p-value of `round(fisher.test(nsub.tab)$p.val,3)`). As discussed in Section XXX, we may not have enough power to detect differences below 5 percentage points. However, we find the same not-significant result in a parametric regression analysis of treatment differences with controls for the demographics and past experience on the platform; see Table **??**. Adding individual covariates reduces variability of outcomes, potentially increasing the power of our test. In particular, Table **??** reports the results from a logistic regression on the probability of making a submissions. Column 1 reports the results from a baseline model with only treatment dummies. Column 2 adds demographics controls, such as the age, education, and gender. Column 3 adds controls for the past experience on the platform. Across all these specifications, the impact of the treatment dummies (including room size) on entry

is not statistically significant.

2.3 Simulation results

List of Figures

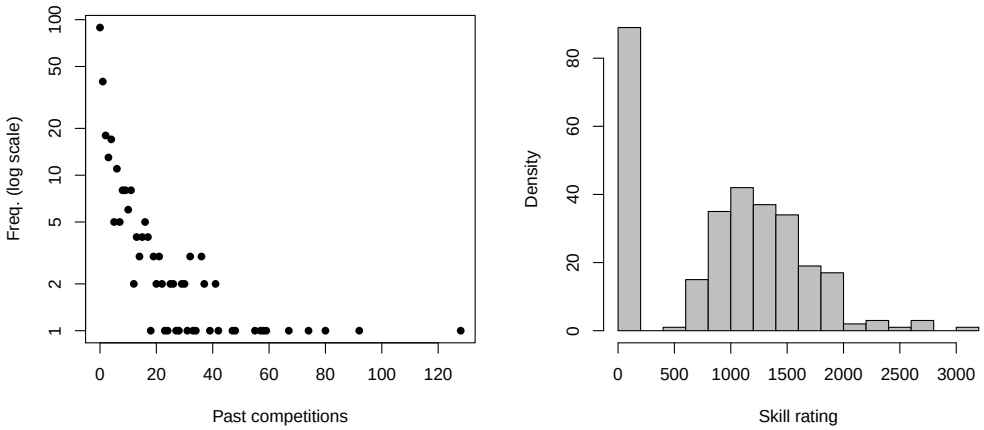


Figure 1: Distribution of the count of past contests (left panel) and the skill ratings (right panel) of the signed-up competitors.

List of Tables

	FALSE	TRUE	Total
Race	73	26	99
Tournament	67	33	100
Reserve	73	27	100
Total	213	86	299