Exploring Repeatability on Mechanical Turk

Kristal Curtis
UC Berkeley
465 Soda Hall
Berkeley, California 94720

Abstract

Repeatability is very desirable.

Introduction

Crowdsourcing has provided numerous opportunities for people in academia and industry to access a numerous pool of workers willing to perform useful work for modest compensation. Many people have taken advantage of crowdsourcing for achieving tasks that are still difficult for computers to complete yet are quite simple for humans, such as image labeling and object characterization. Recently, myriad researchers from computer science as well as disciplines in the social sciences like sociology and economics have turned to Mechanical Turk (MTurk), a popular and flexible crowdsourcing platform, for running human subject experiments.

Due to its large size and impersonal nature, Mechanical Turk gives the illusion of uniformity. However, the reality is that it is incredibly heterogeneous, with participants from around the globe (**?**). In addition, some users of Mechanical Turk (*requesters*, in MTurk parlance) have noted that some workers (ie, *Turkers*) are more active than others (**?**).

A concern that has been raised by both employers and experimenters is the issue of *repeatability*; that is, if a batch of tasks is run under different conditions (eg, different time of day, different day of week), the results may vary. In some cases, this may be a concern because the answer(s) provided by the Turkers may be different, in terms of actual value(s) and/or overall quality. In others, the response time (either time to first answer or time to batch completion) may also be crucial yet highly variable.

In this work, we explore various factors that serve as obstacles to repeatability. We also offer some ideas about how to improve repeatability.

Obstacles to Repeatability

In this section, we investigate the impact of several factors that may serve as obstacles to repeatability.

Zipfian Turker Pool

Some recent studies have shown that for a given group of tasks (ie, a HIT group, where each task is a HIT, or Human Intelligence Task), a small number of Turkers complete a disproportionate amount of the work offered (**?**; **?**). In this work, we will refer to these overly-active Turkers as *super Turkers* ((**?**) refers to them as streakers).

Let us refer to a given HIT group as G. The members of the set $T_S(G_i)$ are the super Turkers who completed HITs for the ith execution of G, where we assume that G is executed n times (ie, at n different occasions). For our purposes, a Turker $t \in T_S(G_i)$ if he/she completes at least k HITs, where $H = |G|$ and $0 < k \leq H$. We will explain how to select k later on.

First, we would like to determine the nature of the intersection between $T_S(G_i)$ and $T_S(G_j)$, where $i \neq j$. It will likely be impacted by the similarity between occasions i and j; ie, if i and j are very different times or dates, you would expect $T_S(G_i) \cap T_S(G_j) = \emptyset$.

Our hypothesis is that for executions G_i and G_j, $i \neq j$, $T_S(G_i) \cap T_S(G_j)$ is small wrt both $|T_S(G_i)|$ and $|T_S(G_j)|$, and that this will cause the results of G_i and G_j to be different.

To validate our hypothesis, we propose the following meta-experiment:

- Given: experiment E, number of occasions n
- Obtain $G_1(E), \ldots, G_n(E)$
- Obtain $T_S(G_1), \ldots, T_S(G_n)$
- Determine the intersections among super Turker sets $T_S(G_i), i \in \{1, \ldots, n\}$.
- Analyze the impact of $T_S(G_i)$ on the results of G_i.

We will explore these ideas in the context of a concrete experiment in the Experiments section.

Contention with Other Tasks

Another factor that could impact repeatability of a HIT group is the *task context*; ie, the number and types of other tasks that are currently live on the MTurk platform. For example, (**?**) observed that even one's own tasks could compete with each other. Therefore, we will also investigate the impact of task context on repeatability. Task context seems more likely to affect response time than actual result values.

In order to measure the number of active HITs on MTurk during the execution of task group G_i, we will scrape the MTurk website once every ten minutes and report the average value observed during the lifetime of G_i, which is defined as the time G_i is posted to the time when all of G_i's tasks have been completed (ie, 100% of each HIT's assignments).

External Events

We also expect that the incidence of external events such as holidays and natural disasters will impact repeatability. In (**?**), for example, the authors note that they had to correct their HIT results that were obtained on a holiday so that they would be comparable with the rest of their results.

Avoiding posting HITs on holidays seems like an obvious workaround; however, the global nature of the MTurk workforce complicates this attempt since Turkers may observe holidays of which the requesters, who are currently all based in the US due to platform restrictions, are unaware.

Unpredictable events that could impact repeatability include natural disasters and unreliable infrastructure, which often disproportionately affect the developing world. Since many Turkers are in India, where they can receive payment in their own currency, this is likely to be an issue.

At this time, we do not attempt to address these issues. However, we do recommend that requesters keep these in mind as potential causes of anomalous results.

Experiments

Zipfian Turker Pool

Contention with Other Tasks

Conclusion