Grant Merrill

Jeremy Schlottmann

1. Which sampling method will your group use (e.g. systematic sample, cluster sample, etc.)?
   * Describe in specific detail how your group will collect data using the sampling method stated. The more detail the better. Just stating that it will be Random will earn no credit on this portion.

We will use a random sampling method. The GitHub API requires an ID as a starting point to collect the next 100 public repositories. We want old and new repositories but not too new. The ID of a repository describes how old a repository is (i.e. a repository with an ID of 1 is the oldest repository). We know that there are somewhere around 60+ million repositories (public and private). We will sample random repositories by requesting random IDs from 1 – 10 million. These repositories will have to meet the requirements specified. They are:

* + - A repository must have 3+ collaborators that push changes
    - A repository should be at least two years old since November 15, 2016
    - A repository must use certain languages (this will be determined after gathering a bit of the data to see which languages are “popular enough.”)

1. What sample size (number of observations) are you planning to obtain from your population?
   * If you are doing a study that involves proportions (yes/no questions) then [this site](http://www.nss.gov.au/nss/home.nsf/pages/Sample+size+calculator) can be used to justify your sample size.
   * If you are doing a study that involves quantitative measurements then [Section 5 of Lesson 10](http://statistics.byuimath.com/index.php?title=Lesson_10:_Inference_for_One_Mean:_Sigma_Known_(Confidence_Interval)) can be used to justify your sample size.

We will be gathering at minimum 100 samples for each language we are using. We will collect more than that if there are little to no issues with gathering and parsing the data.

1. How will applying the options in the previous items provide a representative sample for proper inference to your population?

By having at least 100 samples of each language from anywhere to 1 – 10 million for the ID, this will give us repositories with plenty of history to correctly evaluate the activity of a repository, given the languages used.

1. Will your project involve human subjects (Yes or No)?

No

1. Will you be conducting a survey? (Yes or No)?

No