**INEG 2313: Applied Probability and Statistics for Engineers I**

**Course Project**

**Spring Semester 2013**

Sawyer Anderson

Kanat Bektemirov

**Introduction**

This project studies the most popular content posted to social news and entertainment website reddit.com. Content on reddit is submitted by registered users in the form of either links to content on other websites or of text (“self”) posts. Once a user has submitted a post, other users then proceed to vote the post “up” or “down” (also known as upvotes and downvotes). Upvotes, downvotes, and time since submission are used to rank the posts based on how the user chooses to sort them. For example, a post submitted recently that receives a lot of upvotes will quickly rise to the top on the reddit front page for users who have chosen to sort posts by “hot.” Users are also able post comments and reply to comments relating to each specific post.

In this project, the factors influencing the popularity of posts were studied. It was theorized that the most popular posts might share certain characteristics, such as similar submission times, upvotes, downvotes, number of comments, etc. To investigate, data was collected on sample populations of reddit posts. However, because we were interested not in reddit posts as a whole, but rather in the most popular posts, our sampling was limited to the top 10,000 ranked reddit posts of all time[[1]](#footnote-1) . This top 10,000 post limitation allowed for the use of smaller sample sizes to gather meaningful data – because thousands of posts are submitted to reddit every day, and since most receive only trivial numbers of upvotes, downvotes, and comments, it would have taken much greater sample sizes to gather meaningful data about posts which garner massive attention. With practicality in mind, as well as a desire not to overtax reddit’s servers, it was determined that the top 10,000 posts would be a sufficient population to study.

**Data Collection**

All data for this project was collected in the same manner, and all samples were taken within an hour of each other to ensure that any change in the top 10,000 posts of all time was minimal, if not non-existent, between samples. To collect the data, a Python script was written to pull down and write to file the relevant data. This script made use of the Reddit Application Programming Interface (API), as well as the Python Reddit API Wrapper (PRAW) project, to pull down random[[2]](#footnote-2) samples from the top 10,000 posts. Data collected about each post included the date/time of submission (in Unix timestamp format), whether the post was a self-post or a link, and the number of comments, upvotes, downvotes, and score (upvotes-downvotes). It is important to note that **for each of the six parts of this project, a new sample was taken.** That is, the sample of posts used in part 1 are almost certainly not the same posts as those used in part 2. It is possible that some posts may be used in multiple parts, if the randomization process happened to select the same posts when collecting data for each part. That is, sampling without replacement was used when taking samples in each part (a check was included in the Python script to ensure the same post was not pulled twice), but sampling with replacement was used between parts. Thus, it is possible that some posts may be present in the samples of multiple parts, but individual parts do not contain duplicate posts in their samples.

After parsing, the data was stored in .DATA files for future processing and analysis.

**Data Processing**

Parts 2, 4, and 5 make use of a post’s submission time as a random variable. In its original form, the submission time of a post was formatted as a Unix timestamp – that is, the number of seconds from the beginning of January 1st, 1970, UTC to the time the post was submitted. However, before storing the data locally, it was determined that a more useful and informative piece of data was the time of submission on the day the post was submitted. This data allowed the observation and analysis of trends which may or may not be present in the daily submission time of popular posts. With this in mind, the submission times were converted from Unix timestamp to the number of seconds since the start of the submission day. This conversion was also done using a simple Python script which converted the timestamp into a Python datetime object, and then calculated the submission time as .

Parts 2, 4, and 5 all use this time of submission, rather than the original Unix timestamp.

**Part 1 – Discrete Data**

The first variable studied was the number of comments on posts. More definitively,

This variable was determined to be discrete because reddit considers every comment to be equal in its count, regardless of the length of the comment. A sample size of 100 was required, but a sample size of 500 randomly-selected posts from the top 10,000 posts was chosen due to the relative ease of collecting data and the resulting increase in accuracy of the statistics collected on *C*. Figure 1.1 contains the relevant 1-variable statistics for this sample, and Figure 1.2 shows a histogram of the data.

|  |  |
| --- | --- |
| Mean: | 941.328 |
| Median: | 569.5 |
| Mode: | 251 |
| Min: | 22 |
| Max: | 13654 |
| Range: | 13632 |
| Standard Deviation: | 1255.320499 |

**Figure 1.1** – 1-variable statistics about the number of comments in this sample of 500.

**Figure 1.2** – A histogram showing the relative frequencies of various ranges of number of comments

For information on data collection, see **Data Collection**.

**Part 2 – Continuous Data**

The time of submission was studied as a continuous random variable. Although the precision of measurement was only down to the second, time is continuous, and so is still treated as a continuous random variable. See **Data Processing** for why it was determined to use daily submission time as the random variable:

Figure 2.1 lists the relevant 1-variable statistics from this sample of 500 posts, and Figure 2.2 shows the relative frequencies of various ranges of submission times.

|  |  |
| --- | --- |
| Mean: | 48211.31 |
| Median: | 52412 |
| Mode: | 61065 |
| Min: | 395 |
| Max: | 86218 |
| Range: | 85823 |
| Standard Deviation | 23877.86 |

**Figure 2.1** – 1-variable statistics from a sample size of 500 on *T*.

**Figure 2.2** – A histogram showing the relative frequencies of a range of submission times

Parts 4 and 5 look more in-depth at *T*.

For information on data collection, see **Data Collection**.

**Part 3 – Are Random Variables Independent?**

It was hypothesized that a dependency might exist between the number of comments on a post and its overall rank. That is, it was thought that posts with a higher score, and thus a higher ranking /position, might be more commented upon than posts with a lower score. For this part, the following random variables were tested for independence:

More simply, *P* can be thought of as the popularity ranking of a post, 1 being the most popular post of all time, with subsequent posts being less popular.

Figure 3.1 shows a scatter plot of the position and number of comments of a random sample of 500 posts from the top 10,000 reddit posts.

**Figure 3.1** – A scatter plot of a random sample (*n*=500) of the position of reddit posts vs. the number of comments on those posts.

A cursory examination of this scatter plot (and a more detailed examination of the plot in the project’s spreadsheet) shows no obvious trend or correlation between position and number of comments. From this data, it can be concluded that the number of comments on a top 10,000 reddit post and that post’s position among all other top 10,000 posts are independent. It is important to note that this conclusion is **NOT** valid for all reddit posts, only for those in the top 10,000.

To satisfy the reader’s curiosity, the major outlier in this sample with 28,360 comments was a post by Microsoft Founder Bill Gates inviting reddit users to ask him any questions they might wish to ask.

For information on data collection, see **Data Collection**.

**Part 4 – Is a Random Variable Normal?**

Once again, the time-of-day that a post was submitted *T* (see **Part 2**) was used, this time to test for normality. To perform this test, a new random sample of 500 top-10,000 posts was collected and sorted by ascending *T*. The cumulative frequency CFi for each was then computed using , as well as the corresponding z-value zi such that . *zi* was then plotted against *T*, resulting in the normal probability plot of Figure 4.1.

**Figure 4.1** – A normal probability plot of *T*

Because the normal probability plot of *T* showed a distinctive linear trend, it was determined that *T* could be validly assumed to be normal. That is, .

For information on data collection, see **Data Collection**.

**Part 5 – Hypothesis Test on a Population Mean**

It is widely believed that top scoring posts on reddit are submitted in the morning with regard to time zones in the United States. A number of plausible explanations exist for this phenomenon, primarily that most reddit users browse reddit first thing in the morning upon arriving at work or class. Whatever the explanation, it was hypothesized that the average submission time for the top 10,000 posts was before 0900 US Central Time. Converting 0900 Central to UTC (the time zone of the data) gives 1400. Further converting this time into seconds yields . Thus, the following null and alternative hypotheses were formulated for this test:

Due to the lack of information concerning the true standard deviation of *T*, a sample size was necessary in order for a normality assumption to be valid (that is, in order to be able to use a Z-test, which was chosen over a t-test because of its greater power). For convenience and to maintain consistency with previous parts of the project, a sample size of was once again chosen. For this test, a confidence level of was used. Statistics of the collected sample are shown in Figure 5.1, as well as the relevant test statistic, P-value, and confidence interval.

|  |  |
| --- | --- |
| Sample Average () | 50,422.91 |
| Sample Standard Deviation () | 23,679.58 |
| Test Statistic () | 0.021634 |
| P-Value () | 0.49137 |
| 95% Lower Confidence Interval  () |  |

**Figure 5.1** – Hypothesis test statistics and values for submission time

This test results in a conclusion of failing to reject the null hypothesis. was less then , was far greater than , and μ0 was well within the confidence interval. Interpreted, this conclusion means that we cannot say that the top scoring reddit posts are submitted after 0900 US Central Time. In fact, we can say with 95% confidence that the average submission time of the top 10,000 reddit posts is before 0900 Central.

**Part 6 – Hypothesis Test on a Population Proportion**

For the final part of this project, the proportion of self-posts (text posts) in the top 10,000 reddit posts was tested. More specifically,

The testers’ own experiences, as well as commonly held opinion on reddit, lead to a belief that 10% or less of the top posts on reddit were self-posts. Thus, the following null and alternative hypotheses were formulated:

It was desired to use a Z-test to evaluate this hypothesis, and so it was necessary to approximate X, the number of self-posts in the sample, as a normal random variable . To do so, two conditions must be met: and . Solving these two equations for *n* with yields and , respectively. Once again, for convenience and to maintain consistency, was chosen. Figure 6.1 shows the relevant statistics for this sample and Z-test on the population proportion:

|  |  |
| --- | --- |
| Sample Successes (X) | 36 |
| Test Statistic () | -2.086996779 |
| P-Value () | 0.981555933 |
| 95% Lower Confidence Interval  () |  |

**Figure 6.1** – Hypothesis test statistics and values for proportion of self-posts in this sample of 500 of the top 10,000 reddit posts

1. <http://www.reddit.com/r/all/top/> [↑](#footnote-ref-1)
2. Selection was randomized using Python’s pseudo-random number generator and a truly random seed generated by random.org. [↑](#footnote-ref-2)