The Refugee Center Website Traffic

Huafeng Zhang

I. Introduction

The Refugee Center is an organization to help refugees from all over the world start their new lives in the united states of America. They provide free online education, community building for refugees to learn and share information with each other.[1] When the website needs maintenance, sometimes it must be taken offline. So the refugee center staff wants to find a day or days with the fewest sessions to make website changes without conflicting with their website visitors. To help them find such a day or days, we need to answer the following questions: Is there a difference in the mean sessions among different days? If so, which day(s) have fewest visitors? After plotting the data (see Figure 1), we found the distribution of Monday sessions is heavy tailed and the distributions of Thursday and Saturday sessions are right skewed, so we transform the data logarithmically. In order to find a day(s) that has fewest sessions, we will test if there is a difference between the mean of the log(session) on each day and find which day(s) have fewest visitors.

II. Sampling Design & Data Collection

We use Google Analytics to record visitors' behaviors on the Refugee center website, such as the duration of a session, number of sessions in a day, page path of a session and so on. Since we are interested in finding a day or days in a week that has fewest sessions, we will collect the data that includes date and number of sessions in a day. We choose to study data on sessions from May

1st to October 31st in 2016 because it was in May that the traffic of this newly designed website becomes stable.

The refugee center staff's desired precision is that the point estimate of mean sessions of a day in a week is within 2% of the mean sessions in a day and the associated 95% confidence interval is within at most 4% of that mean. Therefore, the mathematical formula for the desired precision is: $P(|\bar{y} - y_U| \leq d) = 1 - \alpha$, where $d=0.02 * y_u$, $\alpha = 0.05$. We know from the mathematical formula for confidence interval, the absolute precision $e=Z_{\alpha/2} \sqrt{(1-\frac{n}{N})}\frac{S}{\sqrt{n}}$. When we solve this we get $n=\frac{n0}{1+(n0/N)}$, where $n_0= (Z_{\alpha/2}S/d)^2$. Since we already have the population data, the variance of the data is $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y} - y_u)^2$. We have N = 183, $y_u = \sum_{i=1}^{N}\frac{yi}{N}$ =281.21, so S^2= 4756.77, $S=\sqrt{4756.77}$ =68.97, d = 0.02 y_u=5.62. Substituting these values into the formula, we get the sample size n =138.97 ≃ 140. (N is the population size, $\bar{y}$ is the mean sessions from our sample data, y_u is the mean of the sessions in the population data, d is the absolute value of maximum allowable difference between $\bar{y}$ and y_u, S^2 is the population variance of the sessions, $Z_{\alpha/2}$ is the Z critical value at significance level α, and y_i is the session on day i from the sample data).

To organize the raw data, we will first add days of the week to the population data; that is, we stratify the data into seven different groups by day of the week. Then based on the sample size we calculated above, we randomly sample 20 days (140/7=20) without replacement from each group.

I choose stratified sampling because the measurements become more manageable when the population is grouped into stratums. Stratified sampling will ensure that estimates can be made with equal accuracy on different days, and that comparisons of days can be made with equal statistical power,[2] which is the probability of rejecting the null hypothesis when the alternative hypothesis is true.[3] Furthermore, stratified sampling only focuses on important subpopulations which is the key point of our questions.

After comparing stratified random sampling to other sampling methods, we find simple random sampling doesn't provide subsamples of the population, but we need to divide the population data into seven groups for the interest of our questions. We also find in quota sampling, "the selection of the sample is nonrandom,"[5] which indicates that not every observation get a chance of selection, but we want to randomly select different days to get less biased samples. Therefore, stratified sampling is the most appropriate method in this study.

III. Statistical Procedures Used

We plot the data (Figure 1) and find that the distribution of the data is right skewed. Computing the variance of sessions on different days yields the following: $VAR_M = 1427.629$, $VAR_T = 3163.292$, $VAR_W = 6557.629$, $VAR_T = 1093.882$, $VAR_F = 1958.661$, $VAR_S = 1632.274$, $VAR_U = 1904.116$. Clearly, the variances of sessions on different days differ widely. So it is appropriate to logarithmically transform the data of sessions on each day to make the skewed distribution of sessions less skewed.

In this study, the explanatory variable is a day of the week: Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday. The response variable is the log(sessions) in a day. Since we want to test if there is a difference among seven different days of the week, moreover, we only have one categorical explanatory variable and one quantitative response, we will use the one-way ANOVA test first to analyze if there is a difference in the mean log(sessions) on different days. The null hypothesis is that there is no difference between the mean sessions among these days (H_0: $\mu_M = \mu_T = \mu_W = \mu_R = \mu_F = \mu_S = \mu_U$, (μ_M denotes the mean logged sessions on Monday, μ_T, μ_W, μ_R , μ_F , μ_S, μ_U are for Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday respectively). My alternative hypothesis is that at least one day is different from other days, i.e. that H_α: Not all μ_M, μ_T, μ_W, μ_R , μ_F , μ_S, μ_U are equal.

The assumptions for our one-way ANOVA test include independent observations, equal variance and normality of the residuals. The assumption of independent observations is met because these observations are randomly sampled from the population, meaning that the sessions on different days will not affect each other. As we can see from the Residuals vs Fitted plot (Figure 2): there does not appear to be a clear curve remaining in the residuals, all the variances of the logarithmically transformed data are now approximately equal ($VAR_M = 0.039$, $VAR_T = 0.019$, $VAR_W = 0.019$,$VAR_T = 0.019$, $VAR_F = 0.023$, $VAR_S = 0.026$, $VAR_U = 0.035$), so the assumption of equal variance is met. By examining the Normal QQ plot (Figure 2), we found that the distribution of Monday sessions is heavy tailed and the distributions of Thursday and Saturday sessions are right skewed, that is, the assumption of normality of the residuals is not met. But the

one-way ANOVA is considered as a robust test against the normality assumption when the sample sizes are equal, in this study they are equal, so we need not worry about this assumption.

If there is evidence of a difference between the mean of log(sessions) among different days, then we will use the Tukey-Kramer procedure to test which day(s) are different from the others by checking all pairwise comparisons of the means. If the 95% confidence interval for a pair of groups does not include zero (i.e. the p-value for testing the difference of the group is less than 0.05) , then there is evidence that the mean of log(sessions) of the two groups are different. This procedure allows us to find the group(s) (i.e. day(s)) that have fewest sessions.

We think the mean of log(sessions) is the best estimation of daily website traffic because log(sessions) data allows us to use the one-way ANOVA test, a parametric test. Even though nonparametric tests don't require the assumption of normality of observations, such tests usually have less statistical power than parametric tests. Therefore, we will be more likely to test for a significant effect when one truly exists if we apply a parametric test. In our case, the parametric test (the one-way ANOVA test) performs well because the assumptions of independence and equal variance are met. Furthermore, all seven sample sizes in this study are larger than 15 and they are equal.[4]

IV. Summary of Statistical Findings

The study provides strong evidence that there is a difference in the mean of log(sessions) between different days (two-sided p-value=0.0001 from a one-way ANOVA F-Test with test statistic F=25.396, degrees of freedom n_1=6, n_2=133).

After running the Tukey-Kramer test, we found that the p-values in the pairwise comparisons of Friday and the other five days (except Sunday) are smaller than 0.01. This is true as well for the p-values in the pairwise comparisons of Saturday and the other five days (except Sunday), and for the p-values in the pairwise comparisons of Sunday and the other four days (except Friday and Saturday) (see Figure 3). And the estimated difference between each of these pairwise comparisons is less than zero. So there is evidence that there are fewer sessions on Friday, Saturday and Sunday compared to the other four days. On the other hand, the p-values of the pairwise comparisons of means other than those discussed above are fairly large, and their 95% confidence intervals contain zero (see Figure 4), which implies that these groups are similar to each other.

Note that the mean of log(sessions) on Sunday is similar to the mean of log(sessions) on Saturday and Friday (see compact letter display: Figure 5). But the mean of log(sessions) on Saturday is different from Friday. It is estimated that the difference between the mean of log(sessions) on Saturday and the mean of log(sessions) on Friday is -0.212, that is, the median of sessions on Saturday is 0.81 ($e^{-0.212}$) times the median of sessions on Friday with an associated 95% confidence interval from 0.69 ($e^{-0.376}$) to 0.95 ($e^{-0.048}$). And there is little to no evidence that

there is difference between the mean of log(sessions) on Saturday and the mean of log(sessions) on Sunday. (two-sided p-value=0.58)

V. Conclusion

This is an observational study, so we cannot infer a causal relationship between days of the week and website traffic. However, since the data are randomly selected from the population, we can make the inference of the study results to the population if the traffic of the website stays stable. If there is a sudden and unpredicted change, for instance if the organization is disbanded or if there is a sudden surge of refugees in the world because of an unanticipated war, we will no longer be able to use this result. Assuming the web traffic remains as stable as it was during the six months we studied, then the refugee center staff should make website changes on Saturday or Sunday to avoid inconveniencing their website visitors.

After this project, I have a better understanding of stratified random sampling, and I learned how to compare it to other sampling schemes such as simple random sampling and quota sampling. I also learned how to pull data from Google Analytics with R. This project was also a good practice for me to do the one-way ANOVA test.

I am surprised that I need to spend so much time on data mining to figure out an interesting and meaningful question and choose a sampling scheme. However, even though data mining is time consuming, this process is very useful because sometimes we don't have access to the data for the entire population; even we do, the data may be expensive to collect. The first step is to find a

sampling method that yields a representative sample of the population. Only then can we be sure that the sampling method will lead to accurate estimations of the population at a reasonable cost.

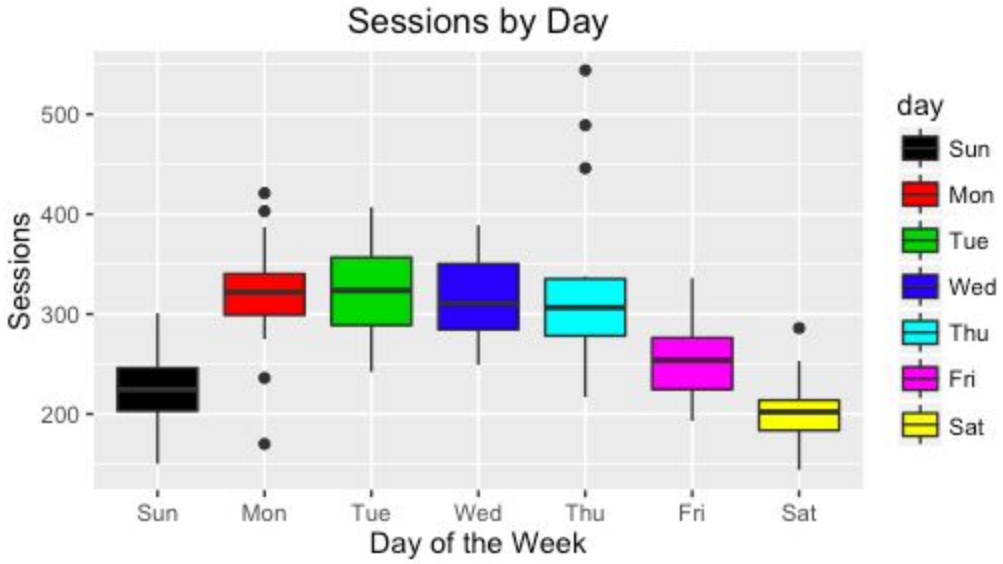
We are required to give a presentation in class to talk about our sampling process and statistical findings. This is no doubt an opportunity for a statistics beginner to learn how to use and talk statistics terms in public. I find it a challenge for me given the fact that English is my second language and statistical terms are strictly defined, so there is a small margin of error allowed.

If I could do the sampling project again without time constraints, I would try to think of more interesting and meaningful questions which require me to apply different sampling methods so that I could practice employing as many sampling methods as possible.

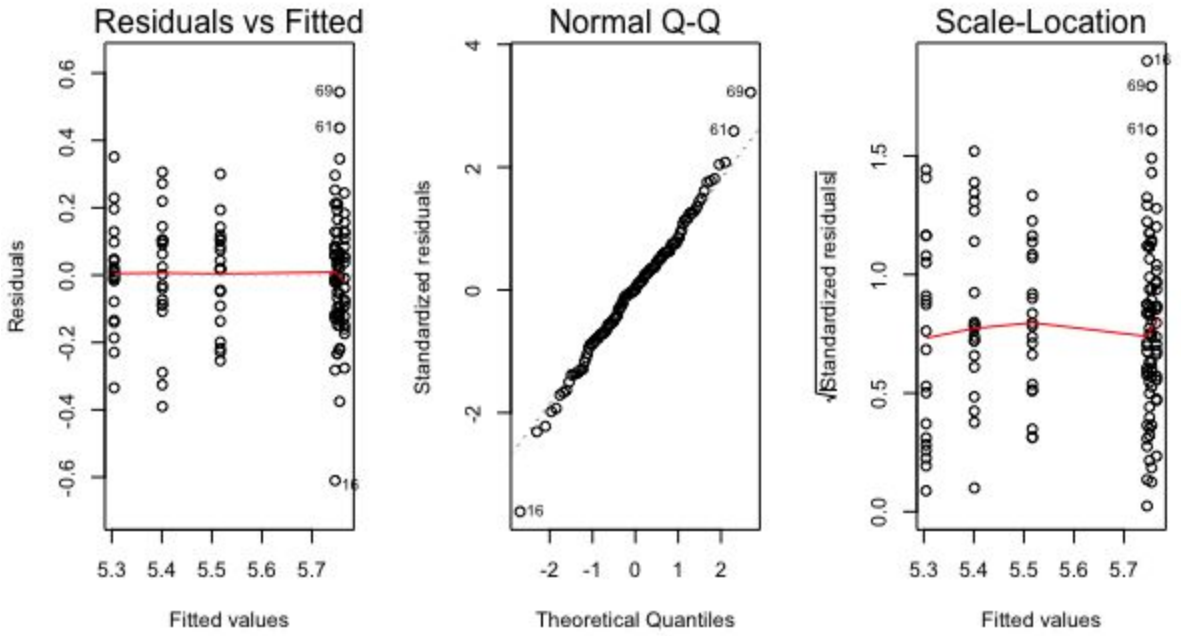
Reference:

[1]. The Refugee Center Online: Help for refugees. (n.d.). Retrieved December 06, 2016, from http://therefugeecenter.org/
[2]. Stratified sampling. (n.d.). Retrieved December 06, 2016, from https://en.wikipedia.org/wiki/Stratified_sampling
[3]. Statistical power. (n.d.). Retrieved December 06, 2016, from https://en.wikipedia.org/wiki/Statistical_power
[4]. Frost, J. (1970). Choosing Between a Nonparametric Test and a Parametric Test. Retrieved December 06, 2016, from http://blog.minitab.com/blog/adventures-in-statistics/choosing-between-a-nonparametric-test-and-a-parametric-test
[5] Sampling (statistics). (n.d.). Retrieved December 06, 2016, from https://en.wikipedia.org/wiki/Sampling_(statistics)

Appendix:



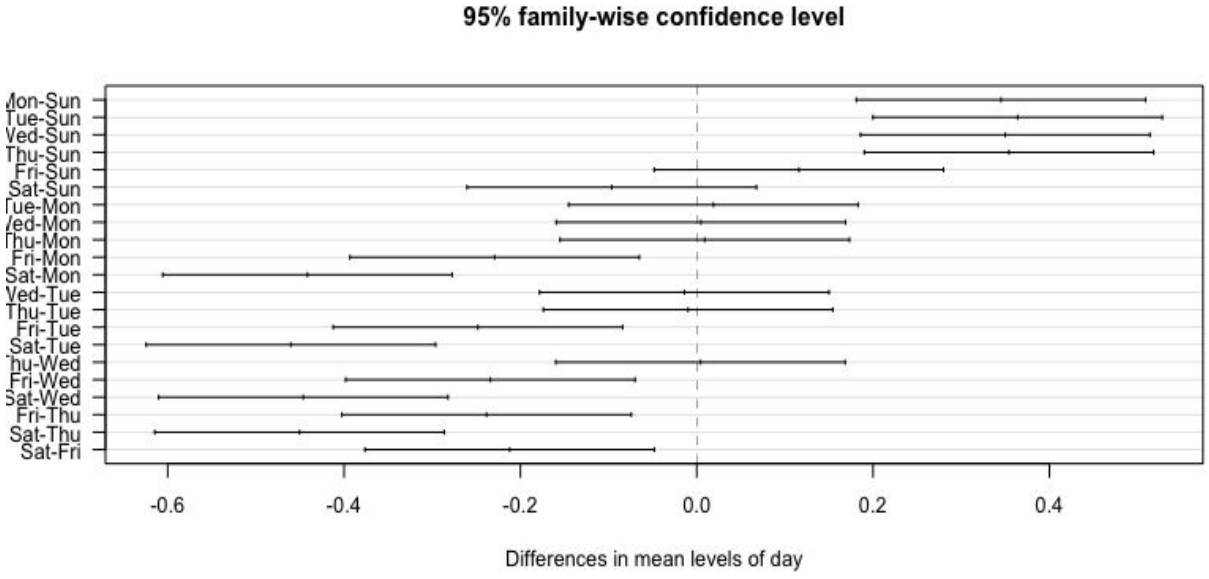
(Figure 1)



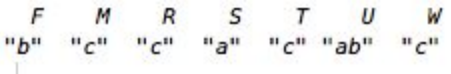
(Figure 2)

	diff	lwr	upr	p adj
Mon-Sun	0.34511485	0.1809828	0.5092469	0.000000
Tue-Sun	0.36401746	0.1998854	0.5281495	0.000000
Wed-Sun	0.34988635	0.1857543	0.5140184	0.000000
Thu-Sun	0.35421249	0.1900805	0.5183445	0.000000
Fri-Sun	0.11584640	-0.0482856	0.2799784	0.351076
Sat-Sun	-0.09636877	-0.2605008	0.0677632	0.578648
Tue-Mon	0.01890261	-0.1452294	0.1830346	0.999862
Wed-Mon	0.00477151	-0.1593605	0.1689035	1.000000
Thu-Mon	0.00909764	-0.1550344	0.1732297	0.999998
Fri-Mon	-0.22926845	-0.3934005	-0.0651364	0.001004
Sat-Mon	-0.44148362	-0.6056156	-0.2773516	0.000000
Wed-Tue	-0.01413111	-0.1782631	0.1500009	0.999975
Thu-Tue	-0.00980497	-0.1739370	0.1543270	0.999997
Fri-Tue	-0.24817106	-0.4123031	-0.0840391	0.000261
Sat-Tue	-0.46038623	-0.6245182	-0.2962542	0.000000
Thu-Wed	0.00432613	-0.1598059	0.1684581	1.000000
Fri-Wed	-0.23403995	-0.3981720	-0.0699079	0.000720
Sat-Wed	-0.44625512	-0.6103871	-0.2821231	0.000000
Fri-Thu	-0.23836609	-0.4024981	-0.0742341	0.000530
Sat-Thu	-0.45058126	-0.6147133	-0.2864492	0.000000
Sat-Fri	-0.21221517	-0.3763472	-0.0480832	0.003142

(Figure 3)



(Figure 4)



(Figure 5)