Go West, (Gluten-Free) Young Man

sdbj2063

November 14, 2015

I Title

"Go West, (Gluten-Free) Young Man"

Gluten-Free Restaurants on Yelp.com Follow the United States Population Migration to the West of the Mississippi River

II Introduction

Our family has followed a gluten-free diet for health reasons for eight years, and my spouse travels extensively on business. These reasons make eating out a serious decision. Finding a reliable resource for locating gluten-free restaurants was a concern. Could **Yelp.com** serve that purpose?

To evaluate the quality of the *Yelp.com academic dataset*, I needed a benchmark issue. One question was whether Yelp.com data followed population trends in the United States. Since the U.S. took its first census in 1790, the population has moved steadily west. *Centers of Population - Geography - U.S. Census Bureau* The 2010 census recorded the mean center of the U.S. population as west, not east, of the Mississippi River. *Position of the Geographic Center of Area, Mean and Median Centers of Population: 2010* The Mississippi River has always been a significant physical and cultural divide for the U.S., as shown in the map below.

Map of Mississippi River | Mississippi River Cruises:

Apparently businesses followed that population trend. For example, a 2013 Huffington Post article cited a study by an online food delivery order platform, *GrubHub*, that ranked U.S. cities in the Top 10 in two categories–one for the highest number of gluten-free delivery orders and a second for the highest number of gluten-free options on the menu. The highest number of cities on both lists were west of the Mississippi River. And **Phoenix, Arizona**, a western city in the Yelp.com data set, was fifth on the list of cities with restaurants offering the most gluten-free menu selections. *"Gluten-Free Takeout: Which Cities Have The Most G-Free Friendly Restaurants?"*

The question was whether the *Yelp.com academic dataset* followed this trend, with more businesses located west of the Mississippi River than east of it. I chose to look at the number of U.S. businesses in the Yelp.com academic challenge `yelp_academic_dataset_business.json` dataset to determine the U.S. business population parameters. Then I sampled that population to determine if the dataset reflected the trend of more gluten-free restaurants west than east of the River.

III Methods and Data

III.A. System Information

The technical configuration for this project consisted of the platform x86_64-w64-mingw32/x64 (64-bit), the operating system Windows 8 x64 (build 9200) and R version 3.2.2 (2015-08-14). The `memory.size()` was 39.5 Mb, and the

`memory.limit()` was 3717 Mb. The machine has an Intel core i5 processor with 8 Gb RAM. I encountered problems in loading `yelp_academic_dataset_review.json` using `knitr` due to application memory constraints for objects, not due to the equipment. **Due to these memory constraints, limited access to external database applications and no access to database-creation R packages that managed large sets of data outside R memory, I limited my focus to business information only.**

III.B. Data

Loading the **JSON** data file consumed the bulk of the processing time.

```
mysystemtime_business <- system.time(mydata_business <- fromJSON(sprintf("[%s]",
    paste(readLines(json_file_business), collapse = ","")), flatten = TRUE))
```

The code assumed that the user downloaded the *Yelp.com* dataset from the repository *Yelp Dataset Challenge*, unzipped the seven files into the working directory, set R to that working directory and set a `memory.limit()` appropriate to handle the files. Using the above system configuration and the `jsonlite` library, loading and flattening the data took an elapsed time of 34.01 seconds for the `yelp_academic_dataset_business.json` file. For the data agreement, please see *Dataset Challenge Academic Dataset Terms of Use*.

The file `yelp_academic_dataset_business.json` contained business information such as an address and key descriptors. To more efficiently process the data, I reduced it to critical fields, resulting in one large dataset.

```
mydata_business <- mydata_business[, c(1, 4, 5, 10, 100)]
mynumber_business <- nrow(mydata_business)
```

The data frame `mydata_business` contained **61184** rows and five fields, including location data fields `city` and `state`. The business data also contained two content-rich fields for business descriptors, `categories` and `attributes`. The latter included the `Dietary Restrictions` subclass with a logical field for `gluten-free`. One of these proved useful for identifying gluten-free businesses.

III.B.1 Exploring U.S. State Data

The first issue was how to segregate businesses by location. The business information contained `city`, `state`, neighborhood, latitude and longitude. For my purposes the latter three variables were too specific. And city was still too granular for my general question. I focused on `state`.

The *Yelp.com* data suggested selecting businesses by `state` and assigning them to either east or west of the Mississippi River was plausible. There are **26** unique state codes listed in the dataset. Filtering using the following U.S. state codes produced the following codes: **PA, NC, SC, WI, IL, MA** in the east and **AZ, OR, MN, CA, NV, WA** in the west.

III.B.2 Exploring Gluten-Free Data

There were two business fields that held the potential for identifying gluten-free businesses, `categories` and `gluten-free Dietary Restrictions attributes`. The former was a free-text field with category key words, while the latter was a logical TRUE/FALSE choice. Creating a **table of frequencies** for `attributes` data and counting the number of `categories` with `Gluten-Free` shed light on the fields' usefulness.

```
mytable_attributes_gf <- table(mydata_business$`attributes.Dietary Restrictions.gluten-free`)
mycategories_gf <- grep("Gluten-Free", mydata_business$categories, value = TRUE)
mytable_attributes_gf
```

```
##
## FALSE  TRUE
##   166     9
```

While the `attributes` field for `gluten-free Dietary Restrictions` sounded promising, the frequency table above revealed that there were only **175** TRUE/FALSE entries out of the **61184** rows for businesses. A more promising field was `categories`, with **81** unique category combinations with the key category `Gluten-Free` out of a total of **8047** unique category combinations.

III.C. Methods

The hypothesis was that in the Yelp.com dataset the U.S. businesses west of the Mississippi River equaled the number of businesses east of the Mississippi River. The alternative hypothesis was that there were more businesses west of the Mississippi River than east. The second question was whether the number of gluten-free restaurants was the same on both sides, or if there were more in the west. To accomplish my research, I selected a population of U.S. businesses and calculated the sizes for two sample sets to divide into east and west groups: one set of businesses and one set of gluten-free businesses.

III.C.1 Selecting a Population

Since my first question applied to businesses in the United States, I selected all businesses with U.S. `state` codes. The number of U.S. businesses was **52849**.

```
mydata_business_east_west <- mydata_business[mydata_business$state %in% mystates_all,
    ]
```

III.C.2 Calculating the Sample Size

The final step was calculating the sample size for U.S. businesses. Research literature regarding selecting an effect suggested **0.2** for a close effect. The **formula** below returned a sample size of 272 businesses.

```
mysample_calc <- pwr.t.test(d = 0.2, sig.level = 0.05, power = 0.95, type = "one.sample",
    alternative = "greater")
```

IV Results

This section selects a sample of U.S. businesses and classifies them west or east of the River. The follow-on task selects gluten-free U.S. businesses and classifies them west or east of the River. The issue is whether to accept or reject the null hypothesis that the number in the west equals the number in the east.

IV.A. General Business Trends East and West of the River

The first step was to collect a random sample of U.S. businesses and group them by their `state` for east and west of the River.

```
set.seed(338)
mydata_business_sample <- mydata_business_east_west[sample(nrow(mydata_business_east_west),
    replace = FALSE, size = mysample_n), ]
mydata_business_west <- mydata_business_sample[mydata_business_sample$state %in%
    mystates_west, ]
mydata_business_east <- mydata_business_sample[mydata_business_sample$state %in%
    mystates_east, ]
```

After collecting the geographic sets, I counted the number of businesses in each group. The sample set results contained businesses from **AZ, NV** in the west and **WI, PA, IL, NC, SC** in the east. They showed that the business dataset followed the population trend, with **221** businesses west of the River and **51** businesses east of the River. I rejected the null hypothesis. The alternative hypothesis was true; there were more businesses west of the River than east.

IV.B. Gluten-Free Restaurant Trends East and West of the River

Identifying the gluten-free restaurant trend required an additional step of narrowing down the U.S. business population to all U.S. businesses with `categories` containing `Gluten-Free`. I recalculated the sample size for this new population, sampled the businesses and divided them by state.

```
mydata_business_gf <- mydata_business_east_west[grep("Gluten-Free", mydata_business_east_west$categories),
    ]
mynumber_total_gf <- nrow(mydata_business_gf)
```

Narrowing the original set of U.S. businesses by the `Gluten-Free` category reduced the total population to **130**. That number was significantly lower than the total U.S. business population of **52849**. It was even lower than the original recommended sample size, **272**. Using the same `pwr.t.test()` function, I recalculated the sample size and adjusted the expectations for `power`, `significance level` and `effect`. The final calculated sample size was **31**, included in the output below.

```
##
##      One-sample t test power calculation
##
##               n = 30.1596
##               d = 0.5
##       sig.level = 0.05
##           power = 0.85
##     alternative = greater
```

Selecting a random sample of gluten-free U.S. businesses, then grouping them into two groups, east and west of the River, was the second step.

```
set.seed(338)
mydata_business_gf_sample <- mydata_business_gf[sample(nrow(mydata_business_gf),
    replace = FALSE, size = mysample_n_gf), ]
mydata_business_gf_west <- mydata_business_gf_sample[mydata_business_gf_sample$state %in%
    mystates_west, ]
mydata_business_gf_east <- mydata_business_gf_sample[mydata_business_gf_sample$state %in%
    mystates_east, ]
```

The result was **25** businesses west of the River and **6** businesses east of the River. The west set contained the states **NV, AZ**, while the east set contained **PA, WI, SC**. I rejected the null hypothesis. The alternative hypothesis was true; there were more businesses west of the River than east.

V Discussion

Based upon the results compared to U.S. population trends, Yelp.com would be a reliable resource for identifying gluten-free restaurants. Moreover, from an academic and market research perspective Yelp.com could be a reliable population research tool.

The sample sets for U.S. businesses and gluten-free restaurants west and east of the Mississippi River matched the U.S. census trend of more people living west of the River than east. I rejected **H0**, concluding that **Ha** was true. The methodology of using `categories` containing `Gluten-Free` combined with Yelp.com's `state` variable proved meaningful. The pie charts below show the division of sample sets for U.S. businesses and gluten-free U.S. restaurants.

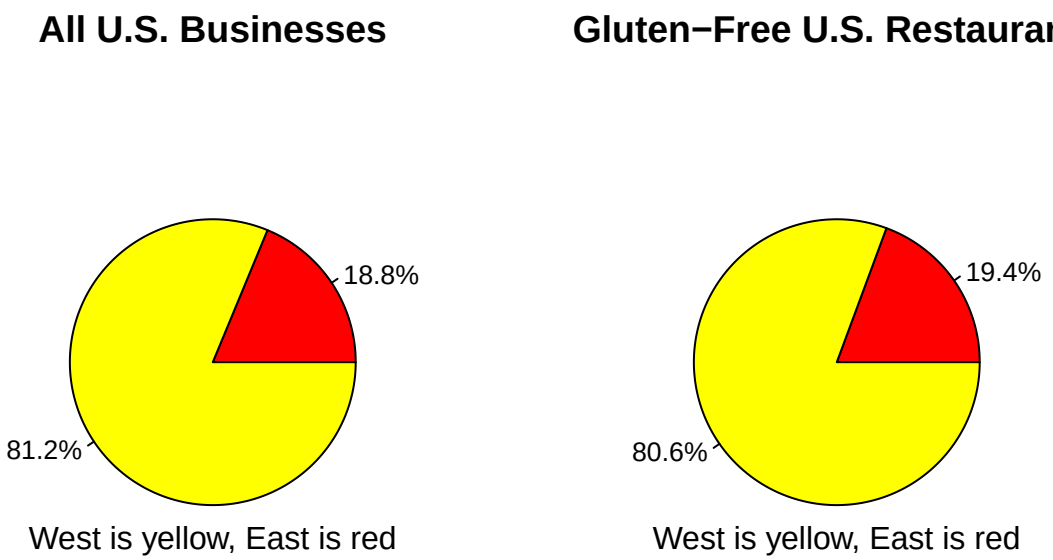

Table of Sample Set Sizes

Variable	All U.S. Businesses	Gluten-Free U.S. Restaurants
States–West	221	25
States–East	51	6
States–Total	272	31

Moving forward, I would examine more data along two paths. First, I would reveiw text that contained "gluten-free" or other permutations. Second, I would expand the search for gluten-free restaurants to likely candidates for gluten-free offerings by looking for other logical, allied `categories`. The table below lists the Top 10 `categories` descriptors that accompanied the terms `Gluten-Free` and `Restaurants`, ranked from highest to lowest frequency. Based upon personal experience there were no surprises, another validation of the Yelp.com data.

```
##    Freq       category
## 47  156    Restaurants
## 31  156    Gluten-Free
## 3    33   Asian Fusion
## 14   32        Chinese
## 45   30          Pizza
## 61   24          Vegan
## 27   24           Food
## 4    12       Bakeries
## 1    12 American (New)
## 62   11     Vegetarian
## 36   11        Italian
## 56    9    Steakhouses
```