ETC5512: Instruction to Open Data

Table of contents

Learning Objectives

- Identify whether data are experimental or observational

- Delineate the data collection methods
- Logically suggest the population that a sample represents

Before your tutorial

Work through the following startR modules:

- Do the module on Projects and Paths (Module 4). *From this week onward we will assume you know how to use RProjects and why these help us organise our analytics work.*
- Do the module on Strategies for troubleshooting R (Module 5).

These should take you ~ 50 minutes.

Package Installation

Ensure you have the packages installed from Week 1's tutorial.

Also install

```
install.packages("tibble")
install.packages("maps")
install.packages("ggthemes")
```

Exercise 1

This question relates to the Tidy Tuesday Data on locations of alternative fuel recharging stations. Have a read through this site, and also visit the link to the data providers, DOT.

a. About the data

Read the details about the data at DOT. How is this data collected, do you think?

The U.S. Department of Energy collects this data in partnership with Clean Cities coalitions and their stakeholders to help fleets and consumers find alternative fueling stations." The implication is that alternative fueling stations provide the details of their service station to the database.

b. Data type

What type of data is this? (observational, experimental, survey, census)

This is closer to a census than any other type. It is beneficial for a fuel provider to be listed in the database, so we would expect that data is available for (almost) all fuel providers.

c. Population vs sample

Describe the population, and what is the sample.

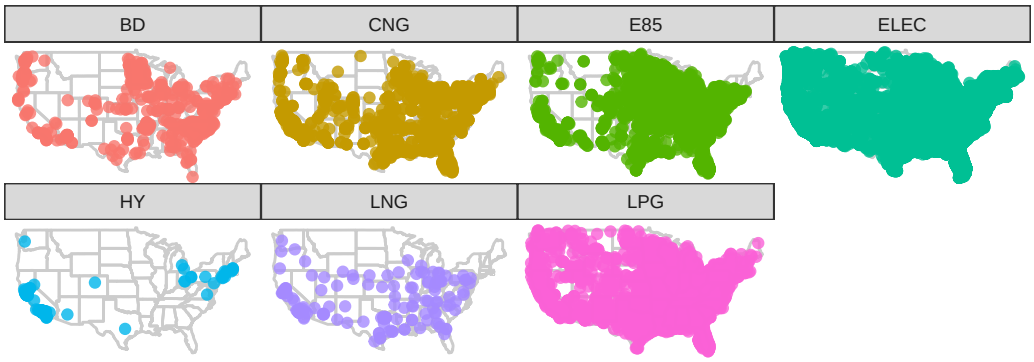
This data could be considered to be the population, not a sample, for a particular time point. It is collected over time, so the data will continue to expand to include new records.

d. Download and plot the data

Download the data and plot the fueling locations on a map, coloured by fuel type.

```
library(tidyverse)
library(ggthemes)
library(maps)
library(mapproj)

# Note you can read data directly from a website
stations <- read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/da

# Get the map data for the USA
usa <- map_data("state")

# Filter to continental USA using map boundary
stations <- stations |>
  filter(between(LONGITUDE, min(usa$long)-1, max(usa$long)+1),
         between(LATITUDE, min(usa$lat)-1, max(usa$lat)+1))

# Plot the sites on a map
# Create a different map for each fuel type

ggplot() +
  geom_path(data=usa, aes(x=long, y=lat, group=group), colour="grey80") +
  geom_point(data=stations, aes(x=LONGITUDE,
                                y=LATITUDE,
                                colour=FUEL_TYPE_CODE),
```

```
            alpha=0.8) +
  facet_wrap(~FUEL_TYPE_CODE, ncol=4) +
  coord_map() +
  theme_map() +
  theme(legend.position = "none")
```

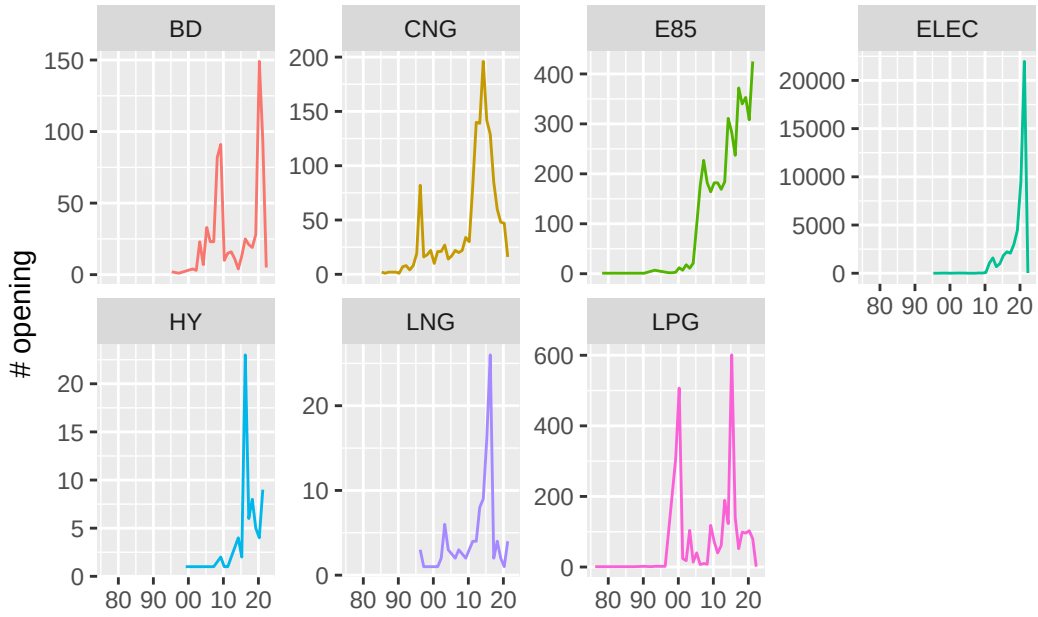


e. New Fuel Stations

Count the number of new stations by month, and make a time series plot by fuel type.

```
# Time line of opening
stations |>
  mutate(OPEN_DATE = as.Date(OPEN_DATE)) |>
  filter(!is.na(OPEN_DATE)) |>
  mutate(m = month(OPEN_DATE),
         yr = year(OPEN_DATE)) |>
  mutate(open_yrmth = as.Date(paste(yr, m, "01", sep="-"), "%Y-%M-%d")) |>
  group_by(open_yrmth, FUEL_TYPE_CODE) |>
  summarise(nopen = n(), .groups = "drop") |>
ggplot(aes(x=open_yrmth,
           y=nopen,
```

```
          colour=FUEL_TYPE_CODE)) +
geom_line() +
facet_wrap(~FUEL_TYPE_CODE, ncol=4, scales="free_y") +
ylab("# opening") +
scale_x_date("", date_labels="%y") +
theme(legend.position = "none")
```



g. Fuel growth

If the question to answer is "which alternative fuel vehicle is the fastest growing?" what is the explanatory (independent, predictor) variable and what is the response variable?

The response variable will be the count of new stations (or cumulative count or rate of growth) relative to time, and the explanatory variable is fuel type. The data suggests that electric is the fastest growing, and it is rapidly expanding.

Exercise 2

Here we will look at the Chocolate bar ratings. Details (brief) of how the data was collected are provided here and more about the data itself is here.

```
chocolate <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/r
```

a. Type of data

What type of data is this? (observational, experimental, survey, census)

Observational

b. Data collection

How is the data collected?

"The Manhattan Chocolate Society's Brady Brelinski has reviewed 2,500+ bars of craft chocolate since 2006, and compiles his findings into a copy-paste-able table that lists each bar's manufacturer, bean origin, percent cocoa, ingredients, review notes, and numerical rating. Related: Craft chocolate makers in the US and Canada, also compiled by Brelinski."

These are ratings given for chocolates, probably sourced and chosen by one person, over a period of time.

c. Population vs sample

Describe the population.

This is a tough one. We'd like to think that these ratings might be useful for our own tastes. However, only one person did the tasting. This likely yields more comparable ratings than if many people tasted them all. Technically the population is that one person, and their perception of the chocolates. Thus the ratings may not infer how other people perceive the chocolates.

d. Response and predictor variables

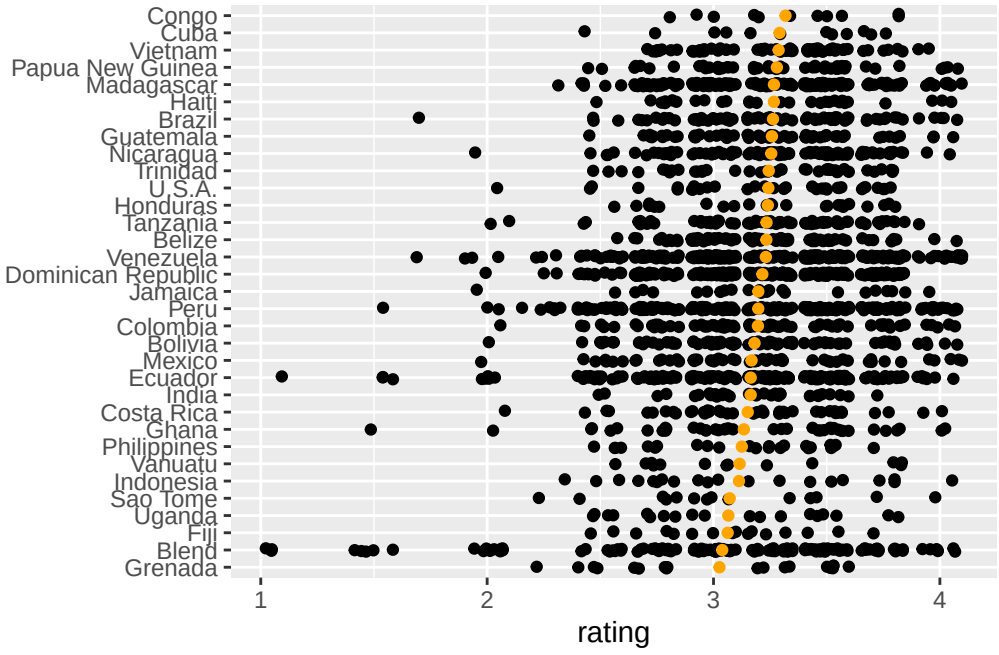
For the question "Which country of origin of the bean obtains the best rating?" state the response and predictor variables.

The response (dependent) variable is rating, and country of bean origin is the predictor (or explanatory or independent variable).

e. Visualise your data

Make a plot to answer the previous question. (Only use countries with more than 10 records.)

```
# Which countries have more than 10
keep <- chocolate |>
  count(country_of_bean_origin, sort = TRUE) |>
  filter(n>10) |>
  pull(country_of_bean_origin)

# Filter to those countries
chocolate_countries = chocolate |>
  filter(country_of_bean_origin %in% keep)

# Plot the results
  ggplot(data = chocolate_countries, aes(x=fct_reorder(country_of_bean_origin, rating, mean)
              y=rating)) +
    geom_jitter(width=0.1) +
    stat_summary(fun = mean, fun.min = median, fun.max = median,
                 geom = "point", colour = "orange") +
    xlab("") +
    coord_flip()
```



There is a lot of variability from country to country, and some chocolates have received very low ratings. Overall Congo has the highest average rating by Brelinski, followed by Cuba, Vietnam, and our near neighbour Papua New Guinea.

In your own time

Read the description of the study titled "Clearing the Fog: Is Hydroxychloroquine Effective in Reducing COVID-19 Progression (COVID-19)".

a. Data type

What type of data is this? (observational, experimental, survey, census)

This is clearly data from a designed experiment.

b. Study Participants

How many subjects participated in the study at the start, and to completion?

180 started in the control group and 360 in the HCQ group. These dropped to 151 and 349, respectively. Thus, 540 subjects started, and 500 completed the experiment.

c. Study details

What are the:

- experimental units?
- factor?
- blocking factors?
- response variable (outcome measures)?

The experimental units are the subjects. The factor is the type of treatment received (control or HCQ).

"Randomization rules were designed by Dr. Wasim Alamgir together with principal investigators and implemented by an independent statistician who was not involved in data analysis. Stratified random sampling was applied to stratify all eligible patients according to age, gender and comorbidities." Participants were assigned to treatment groups by stratified sampling using age, gender and comorbidities as the strata, so these would be the blocks.

"After start of treatment, development of fever $>$ 101 F for $>$ 72 hours, shortness of breath by minimal exertion (10-Step walk test), derangement of basic lab parameters (ALC $<$ 1000 or raised CRP) or appearance of infiltrates on CXR during course of treatment was labeled as

progression irrespective of PCR status." The response variable was whether these symptoms persisted at 5 days.

d. Study desing

How are subjects assigned to treatment groups?

"Randomization rules were designed by Dr. Wasim Alamgir together with principal investigators and implemented by an independent statistician who was not involved in data analysis. Stratified random sampling was applied to stratify all eligible patients according to age, gender and comorbidities." Participants were assigned to treatment groups by stratified sampling using age, gender and comorbidities as the strata.

e. Study results

What were the results of the study?

There was no difference between the subjects treated with HCQ relative to the control. Interestingly, and something of an aside to the main results, is that most patients did not have a progression (see next question, and the numbers reported).

f. Analyse results

Construct the data from the results reported. Compute the proportion of subjects with progression of COVID after 5 days, for the two treatments. Include the standard error of the estimate.

```
hcq <- tibble(trt = c("standard", "standard", "hcq", "hcq"),
              progression = c("all", "yes", "all", "yes"),
              count = c(151, 5, 349, 11))
hcq |>
  pivot_wider(names_from = "progression", values_from = "count") |>
  mutate(p = yes/all) |>
  mutate(se = sqrt(p*(1-p)/all))
```

```
# A tibble: 2 x 5
  trt        all   yes      p      se
  <chr>    <dbl> <dbl>  <dbl>   <dbl>
1 standard   151     5 0.0331 0.0146
2 hcq        349    11 0.0315 0.00935
```

g. Study conculsions

Based on the proportions and their standard errors, why would the result of the study be that HCQ does NOT improve the outcomes of COVID patients?

The proportions for each group are really similar, and the standard error would place both estimates within one standard error of the other. This is strong evidence that both groups of subjects have similar outcomes.

h. Population vs sample

What is the population for this experiment?

This is tough. It is a designed experiment, where age, gender and comobidities have been controlled to be equal between the two groups. The experiment was conducted in Pakistan, with the local population and local conditions. The results may be more applicable for these conditions. However, human DNA is very much the same across all sorts of demographics. That experiments are conducted in various communities is usually considered to be irrelevant, and the results are expected to be relevant broadly to other communities. So the population for this study would be considered to be adult humans, quite generally.