Social Sciences Intro to Statistics

Pset 2: Due MONTH, DATE, YEAR at 11:59pm

Belle Lee

2024-06-15

Overview

Welcome to your second pset of the course. This problem set is intended to give you some practice becoming familiar with descriptive statistics. In this problem set, we are asking you to: create an R project, render your file, load and investigate an R data frame that is stored on the web, and apply some basic functions to better understand distributions.

- Note: Change the values of the YAML header above to your name and the date.

Question 1: Creating an R project

Create an R project

- Create a folder where you want to save files associated with problem set 2. Let's call that folder "problemset2", but you can name it whatever you want.
  - For instance, it could be SSS » problem_sets » problemset2.
- In RStudio, click on "File" » "New Project" » "Existing Directory" » "Browse".
- Browse to find and select your problem set 2 folder.
- Click on "Create Project".
  - An R project file has the extension ".Rproj".
  - The name of the file should be "problemset2.Rproj", or whatever you named the folder.

Save this problemset2.qmd file anywhere in the folder named problemset2.

- At the top of this .qmd file, type in your first and last name in the appropriate place in the YAML header (e.g. "Belle Lee").

- in the date field of the YAML header, insert the date within quotations (any date format is fine).
- Now click the “Render” button near the top of your RStudio window (icon with blue arrow sign) or drop down “File” and select “Render Document”.

  - Alternatively you can use the shortcut: **Cmd/Ctrl + Shift + k**.
  - *Note*: One goal of this assignment is to make sure you are able to render without running into errors.

Question 2: Descriptive Statistics

1. Load the package(s) we will use today: tidyverse

If package not yet installed, then must install before you load. Install in “console” rather than .qmd file

- Generic syntax: `install.packages("package_name")`
- Install “tidyverse”: `install.packages("tidyverse")`

Note: when we load package, name of package is not in quotes; but when we install package, name of package is in quotes:

- `install.packages("tidyverse")`
- `library(tidyverse)`

2. This question asks you to load a dataframe by specifying the `read_csv`()‘ function.

- Url link for data frame: https://raw.githubusercontent.com/bcl96/Social-Sciences-Stats/main/data/netflix_data/Netflix%20Movies%20Dataset%20All.csv

Load the dataframe within this code chunk below.

```
# ANSWER KEY
library(tidyverse)

#load netflix data
netflix_data <- read_csv("https://raw.githubusercontent.com/bcl96/Social-Sciences-Stats/main/
```

2. Find the median for the number of votes. Also define in your own words what does the median represent.

```
# ANSWER KEY
# median(netflix_data$NUMBER_OF_VOTES)

# The median is 41,942.5 votes. Median is the value that is in the middle when you arrange th
```

3. Find the mean for the number of votes. Explain what the mean is and its difference from the median.

```
# ANSWER KEY
# mean(netflix_data$NUMBER_OF_VOTES)

# The mean is 101,966.8 votes. Mean is the average of the dataset, it is the sum of all the v
```

4. Find the standard deviation for the number of votes. In general, what high standard deviation mean compared to a low standard deviation?

```
# ANSWER KEY
# sd(netflix_data$NUMBER_OF_VOTES)

# The standard deviation is 168,874.6, which is quite a high standard deviation. Usually, a h
```

Question 4: Central Limit Theorem

1. Explain in your own words what is the Central Limit Theorem. **ANSWER:**

ANSWER KEY: The Central Limit Theorem states that as the sample size increases, the distribution of the sample mean approaches a normal distribution. This holds true regardless of the shape of the original distribution.

2. Load the ipeds data below from the course website.

```
library(tidyverse)
library(ggplot2)
library(labelled)
library(patchwork)

# Load ipeds dataset from course website
load(url('https://raw.githubusercontent.com/bcl96/Social-Sciences-Stats/main/data/ipeds/outpu

# Create ipeds data frame with fewer variables/observations
df_ipeds_pop <- panel_data %>%
  # keep data from fall 2022
  filter(year == 2022) %>%
  # which universities to keep:
    # 2015 carnegie classification: keep research universities (15,16,17) and master's univer
  filter(c15basic %in% c(15,16,17,18,19,20)) %>%
  # which variables to keep
  select(instnm,unitid,opeid6,opeid,control,c15basic,stabbr,city,zip,locale,obereg, # basic i
         tuition6,fee6,tuition7,fee7, # avg tuition and fees for full-time grad, in-state and
         isprof3,ispfee3,osprof3,ospfee3, # avg tuition and fees for MD, in-state and out-of
         isprof9,ispfee9,osprof9,ospfee9, # avg tuition and fees for Law, in-state and out-of
         chg4ay3,chg7ay3,chg8ay3) %>% # [undergraduate] books+supplies; off-campus (not with
  # rename variables; syntax <new_name> = <old_name>
  rename(region = obereg, # revion
         tuit_grad_res = tuition6, fee_grad_res = fee6, tuit_grad_nres = tuition7, fee_grad_n
         tuit_md_res = isprof3, fee_md_res = ispfee3, tuit_md_nres = osprof3, fee_md_nres = o
         tuit_law_res = isprof9, fee_law_res = ispfee9, tuit_law_nres = osprof9, fee_law_nres
         books_supplies = chg4ay3, roomboard_off = chg7ay3, oth_expense_off = chg8ay3) %>% #
  # create measures of tuition+fees
  mutate(
    tuitfee_grad_res = tuit_grad_res + fee_grad_res, # graduate, state resident
    tuitfee_grad_nres = tuit_grad_nres + fee_grad_nres, # graduate, non-resident
    tuitfee_md_res = tuit_md_res + fee_md_res, # MD, state resident
    tuitfee_md_nres = tuit_md_nres + fee_md_nres, # MD, non-resident
    tuitfee_law_res = tuit_law_res + fee_law_res, # Law, state resident
    tuitfee_law_nres = tuit_law_nres + fee_law_nres) %>% # Law, non-resident
  # create measures of cost-of-attendance (COA) as the sum of tuition, fees, book, living exp
  mutate(
    coa_grad_res = tuit_grad_res + fee_grad_res + books_supplies + roomboard_off + oth_expens
    coa_grad_nres = tuit_grad_nres + fee_grad_nres + books_supplies + roomboard_off + oth_exp
    coa_md_res = tuit_md_res + fee_md_res + books_supplies + roomboard_off + oth_expense_off
    coa_md_nres = tuit_md_nres + fee_md_nres + books_supplies + roomboard_off + oth_expense_o
    coa_law_res = tuit_law_res + fee_law_res + books_supplies + roomboard_off + oth_expense_o
    coa_law_nres = tuit_law_nres + fee_law_nres + books_supplies + roomboard_off + oth_expens
  # keep only observations that have non-missing values for the variable coa_grad_res
    # this does cause us to lose some interesting universities, but doing this will eliminate
  filter(!is.na(coa_grad_res))

# Add variable labels to the tuit+fees variables and coa variables
  # tuition + fees variables
    var_label(df_ipeds_pop[['tuitfee_grad_res']]) <- 'graduate, full-time, resident; avg tuit
    var_label(df_ipeds_pop[['tuitfee_grad_nres']]) <- 'graduate, full-time, non-resident; avg
    var_label(df_ipeds_pop[['tuitfee_md_res']]) <- 'MD, full-time, state resident; avg tuitio
    var_label(df_ipeds_pop[['tuitfee_md_nres']]) <- 'MD, full-time, non-resident; avg tuition
    var_label(df_ipeds_pop[['tuitfee_law_res']]) <- 'Law, full-time, state resident; avg tui
    var_label(df_ipeds_pop[['tuitfee_law_nres']]) <- 'Law, full-time, non-resident; avg tuit

  # COA variables
    var_label(df_ipeds_pop[['coa_grad_res']]) <- 'graduate, full-time, state resident COA; ==
    var_label(df_ipeds_pop[['coa_grad_nres']]) <- 'graduate, full-time, non-resident COA; ==
    var_label(df_ipeds_pop[['coa_md_res']]) <- 'MD, full-time, state resident COA; == tuition
    var_label(df_ipeds_pop[['coa_md_nres']]) <- 'MD, full-time, non-resident COA; == tuition
    var_label(df_ipeds_pop[['coa_law_res']]) <- 'Law, full-time, state resident COA; == tuit
    var_label(df_ipeds_pop[['coa_law_nres']]) <- 'Law, full-time, non-resident COA; == tuitio

df_ipeds_pop %>% glimpse()
#> Rows: 965
#> Columns: 38
#> $ instnm            <chr> "Alabama A & M University", "University of Alabama a~
#> $ unitid            <dbl> 100654, 100663, 100706, 100724, 100751, 100830, 1008~
#> $ opeid6            <chr> "001002", "001052", "001055", "001005", "001051", "0~
#> $ opeid             <chr> "00100200", "00105200", "00105500", "00100500", "001~
#> $ control           <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 2, ~
#> $ c15basic          <dbl+lbl> 18, 15, 16, 19, 16, 18, 16, 20, 18, 18, 19, 18, ~
#> $ stabbr            <chr+lbl> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", ~
#> $ city              <chr> "Normal", "Birmingham", "Huntsville", "Montgomery", ~
#> $ zip               <chr> "35762", "35294-0110", "35899", "36104-0271", "35487~
#> $ locale            <dbl+lbl> 12, 12, 12, 12, 13, 12, 13, 12, 23, 43, 21, 13, ~
#> $ region            <dbl+lbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ~
#> $ tuit_grad_res     <dbl> 10128, 8424, 10632, 7416, 11100, 7812, 10386, 15325,~
#> $ fee_grad_res      <dbl> 1414, 0, 1054, 2740, 690, 766, 1784, 900, 1000, 190,~
#> $ tuit_grad_nres    <dbl> 20160, 19962, 24430, 14832, 31460, 17550, 31158, 153~
#> $ fee_grad_nres     <dbl> 1414, 0, 1054, 2740, 690, 766, 1784, 900, 1000, 190,~
#> $ tuit_md_res       <dbl> NA, 31198, NA, NA, 31198, NA, NA, NA, NA, NA, NA, NA~
#> $ fee_md_res        <dbl> NA, 3464, NA, NA, 0, NA, NA, NA, NA, NA, NA, NA, NA,~
#> $ tuit_md_nres      <dbl> NA, 62714, NA, NA, 62714, NA, NA, NA, NA, NA, NA, NA~
#> $ fee_md_nres       <dbl> NA, 3464, NA, NA, 0, NA, NA, NA, NA, NA, NA, NA, NA,~
#> $ tuit_law_res      <dbl> NA, NA, NA, NA, 24080, NA, NA, 39000, NA, NA, NA, NA~
#> $ fee_law_res       <dbl> NA, NA, NA, NA, 300, NA, NA, 325, NA, NA, NA, NA, 65~
#> $ tuit_law_nres     <dbl> NA, NA, NA, NA, 44470, NA, NA, 39000, NA, NA, NA, NA~
#> $ fee_law_nres      <dbl> NA, NA, NA, NA, 300, NA, NA, 325, NA, NA, NA, NA, 65~
#> $ books_supplies    <dbl> 1600, 1200, 2416, 1600, 800, 1200, 1200, 1800, 998, ~
#> $ roomboard_off     <dbl> 9520, 14330, 11122, 7320, 14426, 10485, 14998, 8020,~
#> $ oth_expense_off   <dbl> 3090, 6007, 4462, 5130, 4858, 4030, 6028, 4600, 3318~
#> $ tuitfee_grad_res  <dbl> 11542, 8424, 11686, 10156, 11790, 8578, 12170, 16225~
#> $ tuitfee_grad_nres <dbl> 21574, 19962, 25484, 17572, 32150, 18316, 32942, 162~
#> $ tuitfee_md_res   <dbl> NA, 34662, NA, NA, 31198, NA, NA, NA, NA, NA, NA, NA~
#> $ tuitfee_md_nres  <dbl> NA, 66178, NA, NA, 62714, NA, NA, NA, NA, NA, NA, NA~
#> $ tuitfee_law_res  <dbl> NA, NA, NA, NA, 24380, NA, NA, 39325, NA, NA, NA, NA~
#> $ tuitfee_law_nres <dbl> NA, NA, NA, NA, 44770, NA, NA, 39325, NA, NA, NA, NA~
#> $ coa_grad_res     <dbl> 25752, 29961, 29686, 24206, 31874, 24293, 34396, 306~
#> $ coa_grad_nres    <dbl> 35784, 41499, 43484, 31622, 52234, 34031, 55168, 306~
#> $ coa_md_res       <dbl> NA, 56199, NA, NA, 51282, NA, NA, NA, NA, NA, NA, NA~
#> $ coa_md_nres      <dbl> NA, 87715, NA, NA, 82798, NA, NA, NA, NA, NA, NA, NA~
#> $ coa_law_res      <dbl> NA, NA, NA, NA, 44464, NA, NA, 53745, NA, NA, NA, NA~
#> $ coa_law_nres     <dbl> NA, NA, NA, NA, 64854, NA, NA, 53745, NA, NA, NA, NA~
```

3. Show how to find the variance of out-of-state tuition for full-time graduates from the `df_ipeds_pop` dataframe.

```
#ANSWER KEY
# Variance of the out-of-state tuition for full-time graduates in the population
var(df_ipeds_pop$tuit_grad_nres)
#> [1] 94056566
```

4. Show the range of out-of-state tuition for full-time graduates from the `df_ipeds_pop` dataframe. Explain what the min and the max represent.

```
#ANSWER KEY
# Range of the out-of-state tuition for full-time graduates in the population
max(df_ipeds_pop$tuit_grad_nres, na.rm = TRUE) - min(df_ipeds_pop$tuit_grad_nres, na.rm = TRU
#> [1] 62259
range(df_ipeds_pop$tuit_grad_nres) # to see the smallest and largest values in the range
#> [1]  1209 63468
```

Question 5: Distributions and Z scores

1. What does a z-score represent?

ANSWER:

ANSWER KEY: The z-score of an observation is the number of standard deviations away from the mean.

2. Calculate z-score for the out-of-state tuition for full-time graduates from the `df_ipeds_pop` dataframe.

```
# components of z-score
mean(df_ipeds_pop$tuit_grad_nres, na.rm = TRUE)
#> [1] 18839.29
sd(df_ipeds_pop$tuit_grad_nres, na.rm = TRUE)
#> [1] 9698.276

#create new variable z_norm_dist
df_ipeds_pop <- df_ipeds_pop %>% mutate(
  z_tuit_grad_nres = (tuit_grad_nres - mean(tuit_grad_nres, na.rm = TRUE))/sd(tuit_grad_nres
)

#list a few observations
df_ipeds_pop %>% select(tuit_grad_nres,z_tuit_grad_nres)
#> # A tibble: 965 x 2
#>    tuit_grad_nres z_tuit_grad_nres
#>             <dbl>            <dbl>
#>  1          20160           0.136
#>  2          19962           0.116
#>  3          24430           0.576
#>  4          14832          -0.413
#>  5          31460           1.30
#>  6          17550          -0.133
#>  7          31158           1.27
#>  8          15325          -0.362
#>  9          19680           0.0867
#> 10          13356          -0.565
#> # i 955 more rows

# mean of z-score variable
round(mean(df_ipeds_pop$z_tuit_grad_nres, na.rm = TRUE), digits = 4)
#> [1] 0
```

Question 6: Sample Statistics

1. Name 3 examples of sample statistics and define each.

ANSWER:

ANSWER KEY: - Number of cases: The number of data points we see in the sample n (notice that the notations are different from N used for population parameter) - Mean: The average of the sample X - Standard deviation: On average how far each data point is from the mean. It also tells us how well the mean of a variable represents the central tendency of a sample S - Variance: The average of the square of the deviations from the sample's mean value s^2 - Range: The difference between the largest and smallest value in the sample

2. Show what the three sample statistics you selected would be for in-state tution of full time graduates in the `df_ipeds_pop` dataframe.

```r
#ANSWER KEY
# Number of cases
nrow(df_ipeds_pop) # each row is a case
#> [1] 965

# Mean of the cost of book supplies in the population
mean(df_ipeds_pop$tuit_grad_res)
#> [1] 14311.72

# Standard deviation of the cost of book supplies in the population
sd(df_ipeds_pop$tuit_grad_res)
#> [1] 9807.926

# Variance of the cost of book supplies in the population
var(df_ipeds_pop$tuit_grad_res)
#> [1] 96195409

# Range of the cost of book supplies in the population
max(df_ipeds_pop$tuit_grad_res, na.rm = TRUE) - min(df_ipeds_pop$tuit_grad_res, na.rm = TRUE)
#> [1] 62259
range(df_ipeds_pop$tuit_grad_res) # to see the smallest and largest values in the range
#> [1]  1209 63468
```

3. Why is it helpful to have sample statistics when we are trying to understand a population?

ANSWER:

ANSWER KEY: Since it is not always possible to collect information or data on the entire population. We rely on the data collected from the samples to make inferences about the population.

Render to pdf and submit problem set

Render to pdf by clicking the "Render" button near the top of your RStudio window (icon with blue arrow) or drop down "File" and select "Render to PDF"

- Go to the [class website] (Need to fill in classwebsite) and under the “Readings & Assignments” » “Week 1” tab, click on the “Problem set 1 submission link”

- Submit both .qmd and pdf files

- Use this naming convention “lastname_firstname_ps#” for your .qmd and pdf files (e.g. lee_belle_ps1.qmd & lee_belle_ps1.pdf)