INFO370 Problem Set 3: CLT, Descriptive Statistics

January 20, 2022

Instructions

This problemset asks you to think about variables and values, do some descriptive statistics, compute 80-20 ratio, and finally explore climate data. It is less demanding than the previous one in terms of doing tricks with pandas but now you need to do some plots. And there is no escape from understanding your data and thinking about it.

- Comment and explain your results! Just numbers with no explanation will not count! Remember: your task is to convince us that you understand, not just to produce correct results!
- Include the question numbers in your solution! You may leave out the text of the question you are answering.
- Ensure your submission is readable. Depending of the complexity of your code and the choice of variable names you may need more or less explanations. For instance, if you are asked to find largest income, then code

  ```python
  print(largestIncome)
  ```

  needs no additional explanations. But if you choose to call the variable "maxy", then you may need to add a comment:

  ```python
  print(maxy)  # 'maxy' is the largest income
  ```

- Make sure your solutions are your own! It is all well to work together, to talk to other students, help them and let them to help you. But at the end you have to understand their suggestions and write your own solution. Please list the other students you worked together with–this helps to avoid too many red flags when graders find solutions surprisingly similar.

Good luck!

List your collaborators here:

1. ...
2. ...

1 Measures, values and relationships (30pt)

In this (and the next) question you work with "states" data (available on canvas). This describes various data for 50 US states, collected in 1970s. The variables are

Population : population estimate as of July 1, 1975

Income : per capita income (1974)

Illiteracy : illiteracy (1970, percent of population)

Life Exp : life expectancy in years (1969-71)

Murder : murder and non-negligent manslaughter rate per 100,000 population (1976)

HS Grad : percent high-school graduates (1970)

Frost : mean number of days with minimum temperature below freezing (1931-1960) in capital or large city

Area : land area in square miles

You may take a look at lecture notes, section 1.1.1 and 1.1.2, for more about measures and values.

1. (3pt) Consider all variables in data. List their measure type (nominar/ordinal/interval/ratio).
2. (3pt) Explain their valid value range (continous, integer, continous within certain limits...).
3. (4pt) Check if all values are within the range you specified. Are there any invalid values?

Next, you task is to analyze relationship between HS graduation rate and income using the same state data.

4. (2pt) Are these variables of a measure type that permit to ask/answer such a question?
5. (3pt) What is your hypothesis: how might these variables be related? What do you think, why might it be like this?
6. (3pt) Make a plot to address your hypothesis. Comment it: does it seem to confirm or not to confirm your thoughts?
7. (3pt) Now let's split the states into two groups: less-educated (HS gradutation rate less than median) and more-educated (HS above median).

   Hint: create such a variable and add it to the data frame.
8. (4pt) Compute the mean income for both of these state groups. Does the result align well with the plot?

   Hint: you can use `groupby`

9. (3pt) Based on your analysis, what do you think: do states have higher income because of more education, or the other way around: better income states can afford more education? Explain your reasoning!

---

2 Explore inequality (30pt)

2.1 Descriptive analysis (10pt)

Now it is time to explore inequality in different kind of data. There are different inequality measures, here we look at 80/20 ratio. This is briefly covered in lecture notes Section 1.2.2 toward the end, and in python notes Section 6.4.

Your task is to analyze inequality in three distributions: citations of research papers; income of labor market program participants (back in 1978); and size of lunar craters. The datasets are as follows:

- Labor market participants: file *treatment.csv*. This contains data about individuals, some of whom did participate in certain training programs. Use the variable *re78*, real income for 1978.
- Citations: file *mag-30k-citations.csv*, citations of 30,000 research papers in Microsoft Academic Graph. The only important variable in the current context is *citations*, how many times the paper has been cited.
- Lunar craters: file *lunar-impact-craters_v08-2015-09.csv.bz2*. Use the variable "`7. Radius [m]`" (note: the variable name contains numbers, space, and brackets!). This is the craters' radius in meters.

Now the detailed tasks:

1. (2pt) Load all three datasets and do basic sanity checks; remove missings values in the variable of interest—you only need the income (and citations and crater radius) variable below.
   (a) How many cases do we have in each data file?
   (b) Do the values of interest look reasonable?
2. (4pt) Show the distribution of all three data in a histogram. As the histogram may not look good, do it in two ways: a) histogram of income (or citations, or crater size) and b) histogram of log income. In order to avoid issues with log of zero you can do $\log(1 + income)$ instead of $\log income$.

   As an extra challenge, try to figure out how to put 6 histograms in a single figure! (This is not necessary though.)
3. (2pt) Look at the histograms and tell-what do you think, which one describes the least unequal distribution, and which one the most unequal distribution?
4. (2pt) Compute sample mean and standard deviation for all three data. Compare these: how much smaller (or larger) is std. dev compared to the mean?

2.2 80-20 ratio (20pt)

One way to describe inequality is to compute the 20/80 ratio. As the famous example tells, 20% of people own 80% of resources. But what are the ratios in data here? Do we have that 10% of craters "own" 90% of "size"? (very unequal) Or maybe 49% of "papers" possess 49% of citations (very little inequality)? The analytic solution does not exist (afaik) though one can solve the ratio numerically. But we go a simpler way and use a loop to figure out an approximate number.

1. (15pt) Compute the 20/80 ratio for all these three distributions.

   You can do it in a following way (but other solutions are ok too):

   (a) Compute the total income in your sample (i.e. sum of all values)

   (b) pick a quantile (say, upper 10%). Find the corresponding income threshold in the sample. You can use `np.percentile`, in this case it would be `np.percentile(x, 90)` for the top-10 pct threshold (this is the same as lower 90th percentile, so thats why "90").

   (c) Find the total income of the top-10 pct by just summing all income values that are larger than the threshold.

   (d) Compute the wealth share of the top-10 pct. Is this more than 90%? If yes then you should look at a smaller top percentage (e.g. 9pct). If not, look for a larger percentage (e.g 11pct).

   (e) In practice, you want to loop over top percentages (e.g. from 1% to 50%) and see where you get close to correct ratio.

   Hint: the answers are approximately 18, 30, 37

2. (5pt) Which distribution is the most unequal one? Which one the most equal one? Does this corresponds to what did you guess based on the visual impression based on the histograms?

---

3 Global temperature over time (40pt)

In this question you will to work with satellite-based global temperature records. There is quite a bit of debate about how satellite records relate to the actual near-ground temperature, here we simply say that we talk about "lower troposphere temperature", whatever it means. You can download the original dataset from University of Alabama, Huntsville http://vortex.nsstc.uah.edu/data/msu/v6.0/tlt/uahncdc_lt_6.0.txt, but rather download from canvas (file *UAH-lower-troposphere-wide.csv.bz2*) where we have done a little bit of cleaning.

The variables are:

year

month month 1..12

the area of measurement: **globe**, **nh** = north hemisphere, **nh_land** = NH land, **nh_ocean** = NH ocean, **sh** = south hemisphere, **trpcs** = tropics, noext = northern areas outside tropics, soext, nopol = northern polar areas, etc.

globe global temperature, deg C deviation from 1991-2020 average.

globe_land global land temperature. There are many others, including hemishperes, ocean, polar areas etc.

Global warming is thought to bring both higher temperatures but also more extreme weather. Can we see this in the data? Your task is to answer two questions:

a) Do we observe a trend in the global temperature over time in this data?

b) Do we observe a trend in the *temporal variability* (i.e. variability over time) of the global temperature in this data?

We base our conclusions on plots and visual inspection only, we do not compute any time trends and confidence values.

Here are the detailed tasks:

1. (5pt) Are these variables of such a measure type that permit to ask/answer such a question?

2. (2pt) Load the data. Perform basic sanity checks.

3. (5pt) Make a simple plot to address the first question–the temperature trend. Which variables do you want to plot? Comment the result: what, if anything, does the figure suggest?

   Hint: you may want to create a variable for time along the lines $\mathit{time} = \mathit{year} + \mathit{month}/12$

4. (6pt) However, for each month we have a single global temperature reading only so we cannot compute the monthly variance across the globe. Instead, let's compute yearly variance, variance of global temperature over months for each year, and make a plot where years are on the horizontal axis and temperature variance on the vertical axis.

   Hint: use groupby by years.

5. (5pt) In order to be consistent, let's do the same with temperature: compute yearly temperature and repeat the plot with yearly averages.

   But what is "yearly temperature"? Do you prefer yearly mean temperature? Or perhaps yearly median? Discuss the advantages/disadvantages of these measures and pick an appropriate measure. You may also display both.

   Hint: Lecture notes Section 1.2.2 "Describing data" discusses mean and median.

6. (6pt) Finally, let's also make similar plots using decades instead of years.

   Hint: create a decade variable using year and integer division `//`.

7. (5pt) In your decadal plot: what do you think about data quality for 1970s and 2020s?

   Hint: how many observations are there?

8. (6pt) Discuss all your plots and state your conclusions: do you see any temperature trend? Do you see any trend in temporal variability? Which plots do you think illustrate your claims in the best way?

How much time did you spend on this PS?