INFO370 Problem Set 3: Descriptive Statistics

Deadline: Oct 30th 5pm

Instructions

This problems set asks you to do descriptive statistics, sampling, and some mathematical statistics. In particular, you are asked to explore a relationship between two variables, and to explore Central Limit Theorem (CLT).

- Comment and explain your results! Only number with no explanation will not count!
- Do not print too much output. To print a few lines of data for illustration is good. To print 1000 lines is carbage.

Good luck!

1 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 how do 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`.

The variables are:

Year

Mo month 1..12

type the area of measurement: Globe, NH = north hemisphere, SH = south hemisphere, Trpcs = tropics, NoExt = northern areas outside tropics, SoExt, NoPol = northern polar areas, etc. There are separate figures for land and sea

temp Temperature, deg C deviation from 1981-2010 average.

Your task is to answer the question: *do we observe a trend in the global temperature over time in this data?* We base our conclusions just on casual observations, sample means and plots, we do not compute any time trends and confidence values.

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

   Hint: read Lecture notes `https://otoomet.bitbucket.io/machineLearning.pdf/` Section 1.1.1 “Measures: Possible Mathematical Operations”

2. (5pt) Load the data. Perform basic sanity checks. Note: the data is *whitespace separated*. you can load it like

   ```
   pd.read_csv("file.csv", delim_whitespace=True)
   ```

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

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

4. (5pt) Now compute yearly temperature averages and repeat the plot with yearly averages.

   Hint: use groupby by years.

5. (5pt) Next, let’s compute decadal averages and make a similar plot. What do you think about the value of the data points for 1970-s and 2020-s?

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

6. (5pt) Compare all three plots: which of these you find the best to answer *this question*?

7. (5pt) Finally, let’s compare northern and southern polar areas (variables NoPol and SoPol). Use the best of your three approaches above and make a (appropriately labeled) plot with two lines, one for north and one for south polar areas.

8. (5pt) State your conclusions: do you see any temperature trend? Do the trends differ for north and south polar regions?

2 Explore Random Variables (60pt)

In this section you will see how does Central Limit Theorem (CLT) work. CLT states that means of random numbers tend to be normally distributed if the sample gets large, and the variance of the mean tends to be $\frac{1}{S}\operatorname{Var} X$ where S is the sample size and X is the random variable, means of which we are analyzing.

CLT, and how variance and mean value change when sample size increases, plays a very important role in computing confidence intervals later.

The problem contains two tasks: work with Pareto-distributed numbers (continuous distribution), and Bernoulli-distributed numbers (discrete distribution).

2.1 Pareto-distributed Random Numbers (30pt)

As the first task, we look at Pareto-distributed random numbers.[1] Pareto is a popular distribution to describe unequal outcomes, such as human income. It has a single parameter α, often called *shape*. Its pdf is given as

$$f(x) = \alpha(1 + x)^{-\alpha-1}, \tag{1}$$

its expected value (mean) is

$$\mathbb{E}\, X = \frac{1}{\alpha - 1}, \quad \alpha > 1 \tag{2}$$

and its variance is

$$\operatorname{Var} X = \frac{\alpha}{(\alpha-1)^2(\alpha-2)}, \quad \alpha > 2. \tag{3}$$

Now let's generate random numbers from this distribution.

1. (1pt) Choose your sample size N. 1000 is a good number.
2. (5pt) Create a vector of N *pareto(10)* random numbers. Make a histogram of those. Comment the shape of the histogram.

   Note: We choose the parameter $\alpha = 10$ as Pareto gets nasty as α gets too small ($\alpha \leqslant 2$). We just want to steer away from those troubles.

   Hint: use `np.random.pareto(10, size)` to create such numbers.

[1] More precisely, we talk here about Pareto-II or Lomax distribution. This is a shifted version of Pareto-I distribution (see wikipedia for details).

3. (5pt) Compute and report mean and variance of the sample you created (just use `np.mean` and `np.var`). Compare these numbers with the theoretical values computed from (2) and (3).

4. (5pt) Now create N *pairs* of random Paretos. For each pair, compute its mean. You should have N means. Make the histogram. How does this look like?

   Hint: while you can do this using loops, it is more useful to create a $\mathsf{N} \times 2$ matrix of random normals, where each row represents one pair. Thereafter your compute means by rows and you have N means.

5. (5pt) Compute and report mean of the pair means, and variance of the means. Compare these numbers with the theoretical values computed from (2) and (3). However, as CLT tells, the variance now should be just 1/2 of what (3) suggests as size of the pairs $\mathsf{S} = 2$.

6. (2pt) Now instead of pairs of random normals, repeat this with 5-tuples of random numbers (i.e. 5 random numbers per one observations). Do you spot any noticeable differences in the histogram?

7. (1pt) Repeat with 25-tuples...

8. (1pt) ... and with 1000-tuples.

9. (5pt) Comment on the tuple size, and the shape of the histogram.

Hint: consult Openintro Statistics 5.1.3 (p 172-178).

2.2 Discrete Random Variables (30pt)

Now we repeat the same exercises with discrete RV-s. We create a RV

$$\mathsf{X} = \begin{cases} -1 & \text{with probability } 0.5 \\ 1 & \text{with probability } 0.5. \end{cases}$$

One way to sample such values is

```
np.random.randint(0,2, size=100)*2 - 1
```

1. (10pt) Calculate the expected value and variance of this random variable.

   Hint: read Openintro Statistics 3.4 (Random variables), in particular 3.4.2 (Variability). I recommend to use the shortcut formula $\operatorname{Var} \mathsf{X} = \mathbb{E}\,\mathsf{X}^2 - (\mathbb{E}\,\mathsf{X})^2$.

2. (10pt) Repeat the same steps as what you did for random normals, just using the discrete RV X instead.

3. (10pt) Explain why do the distribution resemble the normal more and more as we take mean of a large series of individual values.

   Hint: explain what happens when we move from single values $\mathsf{S} = 1$ to pairs $\mathsf{S} = 2$.

How much time did you spend on this PS?

3 Challenge (not graded)

If this task felt too boring for you, here is a more challenging one. Repeat the Pareto-question with $\alpha = 1.5$ (variance does not exist) and $\alpha = 0.5$ (neither variance nor expected value does exist). Explain what you see!