INFO370 Problem Set 3: Descriptive Statistics

2021 年 7 月 14 日

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 numbers with no explanation will not count!
- Do not print too much output. Printing a few lines of data for illustration is good. Printing 1000 lines is carbage.

Good luck!

1 Explore Central Limit Theorem (60pt)

In this section you will see how does Central Limit Theorem (CLT) work. CLT states two things:

1. Means of random numbers tend to be normally distributed if the sample gets large.
2. Variance of the mean tends to be $\frac{1}{\mathsf{S}}\operatorname{Var}\mathsf{X}$ where S is the sample size and X is the random variable we are analyzing.

   (This is actually a direct result of definition of expectation, and independence not really CLT. But CLT is closely related to this result.)

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 Bernoulli-distributed numbers (discrete distribution), and with Pareto-distributed numbers (continuous distribution).

The task is structured in a way that you may want to create a function that takes in sample size S and outputs all needed results, including the histogram. There will be quite a bit of repetitive coding otherwise.

1.1 Discrete Random Variables (36pt)

We start with a distribution that does not look at all normal. We create a RV

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

One way to sample such realizations is

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

1. (5pt) Calculate the expected value and variance of this random variable. Note: these are theoretical values and not related to any samples.

   Hint: read lecture notes 1.3.4 (Expected Value and Variance), and Openintro Statistics 3.4 (Random variables), in particular 3.4.2 (Variability). I recommend to use the shortcut formula $\operatorname{Var} X = \mathbb{E}\, X^2 - (\mathbb{E}\, X)^2$.

2. (1pt) Choose your number of repetitions R. 1000 is a good number but you can also take 10,000 or 100,000 to get smoother histograms.

   Note: number of repetitions R *is not* the same as sample size S here. You will create samples of size S for R times. For instance you create $R = 1000$ times a sample of size $S = 5$. Please understand the difference, it is a fequent source of confusion!

3. (2pt) Create a vector of R random realizations of X. Make a histogram of those. Comment the shape of the histogram.

   Note: in this case we have $R = 1000$ repetitions and samples of size $S = 1$–we look at individual realizations.

4. (4pt) 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 in 1.

5. (3pt) Now create R *pairs* of random realizations of X (i.e. sample size $S = 2$). For each pair, compute its mean. You should have R means. Make the histogram. How does this look like?

   Hint: while you can do this using loops, it is more useful to create a $R \times 2$ matrix of realizations of X, where each row represents one pair. Thereafter you compute means by rows and you have R pair means. See python notes numpy statistical functions for an example.

6. (3pt) Compute and report mean of the R pair means, and variance of the means.

7. (5pt) Compute the expected value and variance of the pair means, i.e. the theoretical concepts. Compare the theoretical values with the sample values above.

   Note that according to CLT tells, the variance of a pair mean should be just 1/2 of what you got above as for pairs $S = 2$.

8. (2pt) Now instead of pairs of random numbers, repeat this with 5-tuples of random numbers (i.e. $S = 5$ random numbers per one repetition, and still $R = 1000$ repetitions in total). Compare the theoretical and sample version of mean and variance of 5-tuples. Are they similar? Do you spot any noticeable differences in the histogram compared to your previous histogram?

9. (2pt) Repeat with 25-tuples...

10. (2pt) ... and with 1000-tuples.

11. (3pt) Comment on the tuple size, and the shape of the histogram.

12. (4pt) Explain why do the histograms resemble normal distribution as S grows.

    In particular, explain what happens when we move from single values $S = 1$ to pairs $S = 2$. Why did two equal peaks turn into a "⩍"-like histogram?

1.2 Pareto-distributed Random Numbers (24pt)

Next, 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) can be compute by just integrating the pdf but we just rely on the formulas:

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

for expected value and

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

for variance. You will encounter pareto-distributed data again and again, so let's get a little familiar with it.

First let's generate random numbers from this distribution.

1. (4pt) Create a vector of R *pareto*(*5*) random numbers. Make a histogram of those. Comment the shape of the histogram.

   Note: We choose the parameter $\alpha = 5$ 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(5, size)` to create such numbers.

2. (3pt) 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).

   Hint: these should be similar.

[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. (3pt) Now create R *pairs* of random Paretos. For each pair, compute its mean. You should have R means. Make the histogram. How does this look like?

4. (4pt) Compute and report mean of the pair means, and variance of the means.

5. (5pt) Compute theoretical mean and variance of pair means using (2), (3), and CLT. Remember, the variance now should be just 1/2 of what (3) suggests as size of the pairs S = 2. Compare these numbers with the sample versions.

   Hint: your experimental results should be similar to the theoretical ones.

6. (1pt) Now repeat this with 5-tuples of random numbers. Do you spot any noticeable differences in the histogram?

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

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

9. (2pt) Comment on the tuple size, and the shape of the histogram.

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

1.3 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!

You will encounter highly inequal distributions in your practice, and it is useful to recognize those on the histogram.

2 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, on the version on canvas we have done a little bit of cleaning.

The variables are:

Year

Mo month 1..12

the area of measurement: **Globe**, **NH** = north hemisphere, **Land** = NH land, **Ocean** = NH ocean, **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 1991-2020 average.

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* 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.

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

   Hint: read Lecture notes http://faculty.washington.edu/otoomet/machineLearning.pdf Section 1.1.1 "Measures: Possible Mathematical Operations"

2. (2pt) 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 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 need a variable for time along the lines $time = year + month/12$

4. (6pt) However, for each month we have a single global temperature reading so we cannot compute the monthly variance. Instead, let's compute yearly variance, 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 http://faculty.washington.edu/otoomet/machineLearning.pdf Section 1.2.2 "Doing descriptive statistics" 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 of 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?