Exercise 5

Due Friday June 18 2021.

Some files are provided that you need below: E5.zip. **You may not write any loops** in your code.

Summary Statistics and Data Exploration

In this question, you will produce some summary statistics for various data sets. Submit a Jupyter Notebook `summary.ipynb` with whatever code you used to produce the values. There is no specific requirement for the format (and if writing loops here makes you happy, then go ahead).

Each of the `data-*.csv` data sets provided contains (x, y) data points. Create a file `summary.txt` and for each data set, give:

- The mean, standard deviation, and range (min and max) of both the x and y variables.
- The correlation coefficient (r) between the x and y variables.
- A one sentence description of what's in the data set: describe the data set to someone and say what you think someone needs to know about it.

Your code does **not** have to produce the `summary.txt` file. We only want to see the code you used to produce the values.

Reddit Weekends

This question uses data derived from the Reddit Comment archive *(http://files.pushshift.io/reddit/)* , which is a collection of every Reddit comment, distributed as 150 GB of compressed JSON.

I have done some aggregation on that so you don't have to: the provided file `reddit-counts.json.gz` contains a count of the number of comments posted daily in each Canadian-province subreddit, and in /r/canada itself. (The values will differ slightly from The Truth: I haven't done the timezones correctly, so there will be some comments categorized incorrectly around midnight: I'm willing to live with that.) Again, the format is gzipped line-by-line JSON. It turns out Pandas (≥0.21) can handle the compression automatically and we don't need to explicitly uncompress:

```
counts = pd.read_json(sys.argv[1], lines=True)
```

The question at hand: **are there a different number of Reddit comments posted on weekdays than on weekends?**

For this question, we will **look only at values** (1) in 2012 and 2013, and (2) in the /r/canada subreddit. Start by creating a DataFrame with the provided data, and separate the weekdays from the weekends. Hint: check for `datetime.date.weekday` *(https://docs.python.org/3/library/datetime.html#datetime.date.weekday)* either 5 or 6.

Create a program `reddit_weekends.py` for this question. Output is described below. Take the input data file on the command line:

```
python3 reddit_weekends.py reddit-counts.json.gz
```

Student's T-Test

Use `scipy.stats` *(https://docs.scipy.org/doc/scipy/reference/stats.html)* to do a T-test on the data to get a p-value. Can you conclude that there are a different number of comments on weekdays compared to weekends?

Try `stats.normaltest` to see if the data is normally-distributed, and `stats.levene` to see if the two data sets have equal variances. Now do you think you can draw a conclusion? (Hint: no. Just to check that we're on the same page: I see a “0.0438” here.)

Fix 1: transforming data might save us.

Have a look at a histogram of the data. You will notice that it's skewed: that's the reason it wasn't normally-distributed in the last part.

Transform the counts so the data doesn't fail the normality test. Likely options for transforms: `np.log`, `np.exp`, `np.sqrt`, `counts**2`. Pick the one of these that comes closest to normal distributions.

[Unless I missed something, none of them will pass the normality test. The best I can get: one variable with normality problems, one okay; no equal-variance problems.]

Fix 2: the Central Limit Theorem might save us.

The central limit theorem says that if our numbers are large enough, and we look at sample means, then the result should be normal. Let's try that: we will combine all weekdays and weekend days **from each year/week** pair and take the mean of their (non-transformed) counts.

Hints: you can get a "year" and "week number" from the first two values returned by `date.isocalendar()` *(https://docs.python.org/3//library/datetime.html#datetime.date.isocalendar)* . This year **and** week number will give you an identifier for the week. Use Pandas to group by that value, and aggregate taking the mean. Note: the year returned by `isocalendar` **isn't always the same** as the date's year (around the new year). Use the year from `isocalendar`, which is correct for this.

Check these values for normality and equal variance. Apply a T-test if it makes sense to do so. (Hint: yay!)

We should note that we're subtly changing the question here. It's now something like "do the number of comments on weekends and weekdays for each week differ?"

Fix 3: a non-parametric test might save us.

The other option we have in our toolkit: a statistical test that doesn't care about the shape of its input as much. The Mann–Whitney U-test *(https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test)* does not assume normally-distributed values, or equal variance.

Perform a U-test on the (original non-transformed, non-aggregated) counts. Note that we should **do a two-sided test** here, which will match the other analyses. Make sure you get the arguments to the function correct.

Again, note that we're subtly changing the question again. If we reach a conclusion because of a U test, it's something like "it's not equally-likely that the larger number of comments occur on weekends vs weekdays."

Output

The provided `reddit_weekends_hint.py` provides a template for the output that will produce a consistent result that we can check easily. Output all of the relevant (and one irrelevant) p-values from the tests you did on the data in the format provided.

Questions

Answer these questions in a file `answers.txt`.

1. Which of the four transforms suggested got you the closest to satisfying the assumptions of a T-test?
2. I gave imprecise English translations of what the by-week test, and the Mann-Whitney test were actually testing. Do the same for the original T-test, and for the transformed data T-test. That is, describe what the conclusion would be if you could reject the null hypothesis in those tests.

3. Of the four approaches, which do you think actually does a better job of getting an answer for the original question: “are there a different number of Reddit comments posted on weekdays than on weekends?” Briefly explain why. (It's not clear to me that there is a single correct answer to this question.)
4. When are more Reddit comments posted in /r/canada, on average weekdays or weekends?

Submitting

Submit your files through CourSys for Exercise 5.

Updated Wed April 07 2021, 09:16 by ggbaker.