FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION

OF HIGHER EDUCATION

ITMO UNIVERSITY

Report on learning practice # 1

Analysis of univariate random variables

Performed by

Valeria Sakovskaya,

Ekaterina Deviatiarova,

Daria Murova

J4134c

St. Petersburg

2022

# 1. Substantiation of Chosen Subsample

This report provides the analysis of the climate weather of surface in Brazil. Based on univariate statistical analysis, it is possible to determine the meteorological seasonality in the south region of Brazil. Such an analysis is important for the population, since it makes it possible to improve the productivity of economic activities, especially crop cultivation and livestock rearing.

The present dataset covers hourly weather data from 623 weather stations and consists of 26 variables, including 21 continuous ones:

* date (yyyy-mm-dd);
* time (hh:00);
* amount of precipitation in millimetres (last hour);
* atmospheric pressure at station level (mb);
* maximum air pressure for the last hour (mb);
* minimum air pressure for the last hour (mb);
* solar radiation ();
* air temperature (instant) (°c);
* dew point temperature (instant) (°c);
* maximum temperature for the last hour (°c);
* minimum temperature for the last hour (°c);
* maximum dew point temperature for the last hour (°c);
* minimum dew point temperature for the last hour (°c);
* maximum relative humid temperature for the last hour (%);
* minimum relative humid temperature for the last hour (%);
* relative humid (% instant);
* wind direction (radius degrees (0-360));
* wind gust in metres per second;
* wind speed in metres per second;
* brazilian geopolitical regions;
* state (province);
* station name (usually city location or nickname);
* station code (inmet number);
* latitude;
* longitude;
* elevation.

# 2. Plotting a Non-Parametric Estimation of PDF in Form of a Histogram and Using Kernel Density Function

The *probability density function* (*pdf*) is a fundamental concept in statistics. Given the pdf of a random variable , probabilities associated with can be easily computed as

Many problems in statistics can be described as using a (random) sample to infer the underlying (unknown) population. Given a sample , the problem of density estimation is to contruct an estimate of based on the sample.

In nonparametric density estimation, we do not restrict the form of with any parametric assumptions (We still need some smoothness assumptions in order to analyze theoretical properties). It is not easy to give a clear definition of what is “nonparametric”, but sometimes it is useful to think nonparametric as “infinite parametric”.

The kernel density estimator (KDE; sometimes called kernel density estimation) is one of the most famous methods for density estimation. Figure 1 shows the KDE and the histogram of the faithful dataset in Python. The blue curve is the density curve estimated by the KDE.

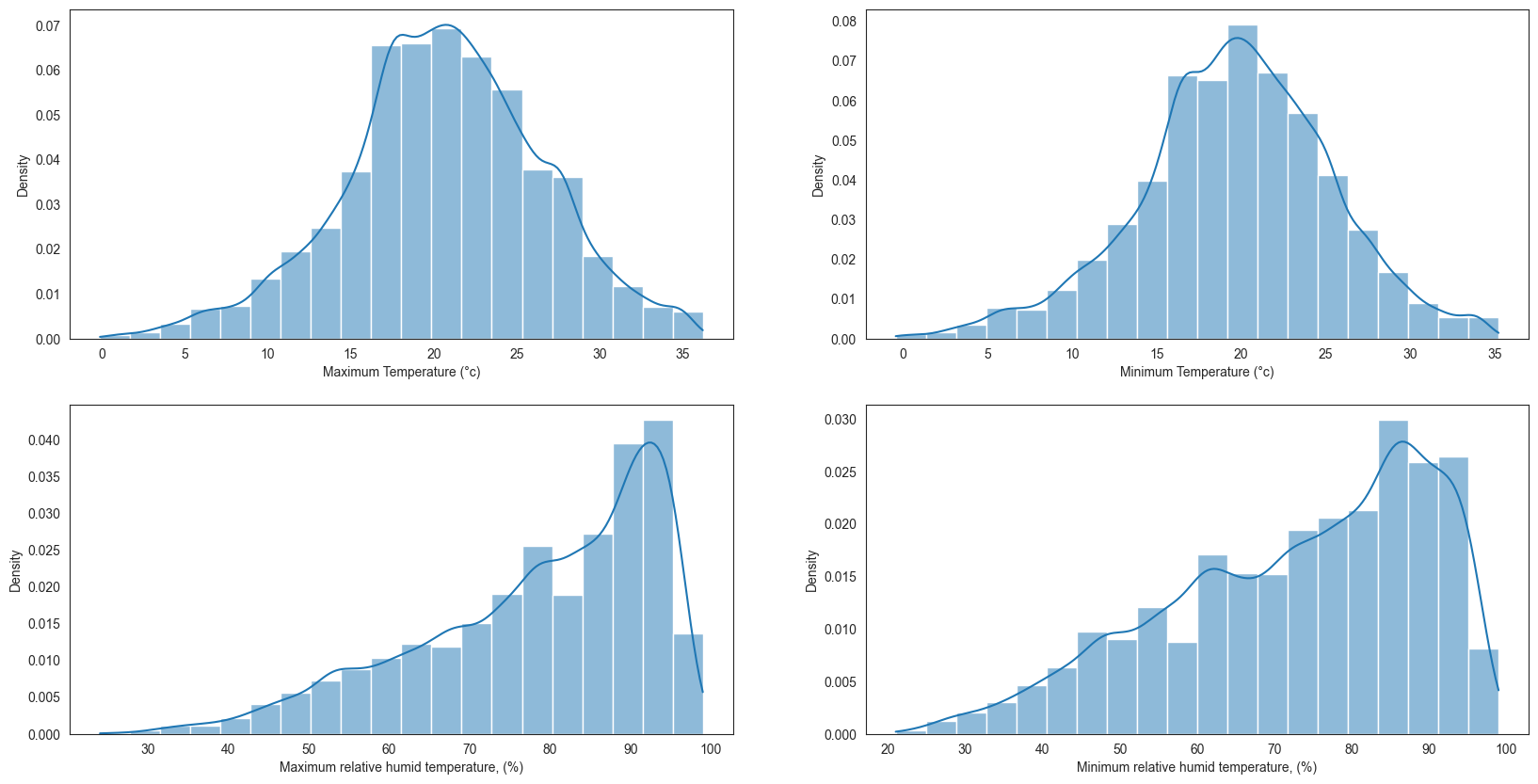


Figure 1 – Histogram and Kernel Density Estimator

Here is the formal definition of the KDE. The KDE is a function

,

where is called the kernel function that is generally a smooth, symmetric function such as a Gaussian and is called the smoothing bandwidth that controls the amount of smoothing. Basically, the KDE smoothes each data point into a small density bump and then sum all these small bumps together to obtain the final density estimate.

As we can see in Figure 1, the median maximum temperature in the south of Brazil varies from 17°c to 23°c.

# 3. Order Statistics Estimation and Its Representation as “Box with Whiskers” Plot

In descriptive statistics, a box plot (also known as box and whisker plot) is a type of chart often used in explanatory data analysis. Box plots visually show the distribution of numerical data and skewness through displaying the data quartiles (or percentiles) and averages.

Box plots show the five-number summary of a set of data: including the minimum score, first (lower) quartile, median, third (upper) quartile, and maximum score.

Construction of a box plot is based around a dataset’s quartiles, or the values that divide the dataset into equal fourths. The first quartile () is greater than 25% of the data and less than the other 75%. The second quartile () sits in the middle, dividing the data in half. is also known as the median. The third quartile () is larger than 75% of the data, and smaller than the remaining 25%. In a box and whiskers plot, the ends of the box and its center line mark the locations of these three quartiles.

The distance between and is known as the interquartile range () and plays a major part in how long the whiskers extending from the box are. Each whisker extends to the furthest data point in each wing that is within 1.5 times the . Any data point further than that distance is considered an outlier, and is marked with a dot.

When a data distribution is symmetric, you can expect the median to be in the exact center of the box: the distance between and should be the same as between and . Outliers should be evenly present on either side of the box. If a distribution is skewed, then the median will not be in the middle of the box, and instead off to the side. It is also possible to find an imbalance in the whisker lengths, where one side is short with no outliers, and the other has a long tail with many more outliers.

|  |  |
| --- | --- |
| (a) Maximum temperature () | (b) Minimum temperature ( |
| (c) Maximum relative humidity temperature (%) | (d) Minimum relative humidity temperature (%) |

Figure 2 – Box plots

The distributions of both maximum and minimum temperatures are approximately symmetric, because both half-boxes are the same length (15). Its interquartile range is  
. Also, the median maximum temperature equals . The minimum value of maximum temperature is , and the maximum is more than . As for both maximum and minimum humidity temperatures’ distributions, we can see that the distribution is skewed, and the median is not in the middle of the box, and instead off to the right side. The imbalance in the whisker lengths can also be found, as the right side is short with no outliers, and the other has a long tail win many more outliers.

Table 1. Percentile Distribution of target variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Quantile** | **0.10** | **0.25 (Q1)** | **0.50 (Q2)** | **0.75 (Q3)** | **0.90 (Q4)** |
| Max temp | 12.9 | 17.0 | 20.6 | 24.6 | 28.1 |
| Min temp | 12.2 | 16.3 | 19.7 | 23.4 | 26.7 |
| Max relative humidity temp | 56.0 | 69.0 | 82.0 | 91.0 | 94.0 |
| Min relative humidity temp | 48.0 | 61.0 | 76.0 | 87.0 | 93.0 |

# 4. Selection of Theoretical Distributions that Best Reflect Empirical Data

Probability distribution fitting is the fitting of a probability distribution to a series of repeated measurements of a variable phenomenon. A distribution with a close fit can be used for various purposes as described in use-cases.

An approach to fit a probability distribution to data is a goodness of fit test. This compares the observed frequency to the expected frequency from the model -hat for any number of classes. In distfit the goodness of fit test is computed with the Sum of Squared Errors (or estimates) (), also named Residual Sum of Squares ().

The describes the deviation predicted from actual empirical values of data. Or in other words, the differences in the estimates. It is a measure of the discrepancy between the data and an estimation model. A small indicates a close fit of the model to the data. is computed by:

where is the -th value of the variable to be predicted, is the -th value of the explanatory variable, and is the predicted value of (also termed -hat).

The theoretic distributions were found using a python package called *distfit*. It is used for probability density fitting across 89 univariate distributions to non-censored data by . The best fitted distribution is returned with the loc, scale, arg parameters which can then be used to compute the probability on new data-points.

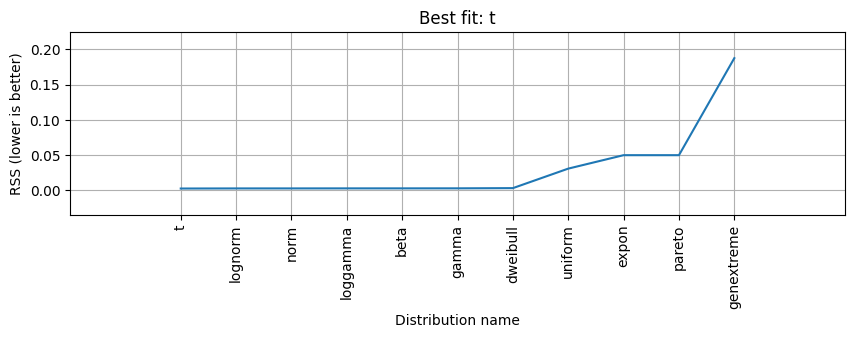


Figure 3 – Summary of evaluated pdfs for the distribution of maximum temperature

To measure the goodness of fit of pdfs, multiple pdfs were evaluated using the RSS metrics. The goodness of fit scores for maximum temperature distribution is presented in Figure 3. The model detected t-distribution as the best pdf but a good score is also detected for the *lognormal*, *normal*, *loggamma*, *beta*, *gamma* and *dweibull* distribution.

The goodness of fit will change according the number of samples that is provided. In this case, more than samples were specified which gave good results.

|  |  |
| --- | --- |
| (a) Maximum temperature | (b) Minimum temperature |
| (c) Maximum relative humidity temperature | (d) Minimum humidity temperature |

Figure 4 – Theoretical distributions

# 5. Estimation of Random Variable Distribution Parameters Using Maximum Likelihood Technique and LS Methods

The least squares and maximum likelihood estimation methods are two commonly used approaches to estimate population parameters from a random sample. MLE indicates how likely the observed sample is as a function of possible parameter values. Therefore, maximizing the likelihood function determines the parameters that are most likely to produce the observed data. From a statistical point of view, MLE is usually recommended for large samples because it is versatile, applicable to most models and different types of data, and produces the most precise estimates.

Least squares estimates are calculated by fitting a regression line to the points from a data set that has the minimal sum of the deviations squared (least square error). In reliability analysis, the line and the data are plotted on a probability plot.

Table 2. Parameters estimated using Maximum Likelihood and Least Squares Methods

|  |  |  |
| --- | --- | --- |
| **Target Variable’s Distribution** | **MLE** | **LS** |
| Student’s t-distribution |  |  |
| Student’s t-distribution |  |  |
| Log-Gamma distribution |  |  |
| Beta distribution |  |  |

|  |  |
| --- | --- |
| (a) | (b) |

Figure 5 – Student’s t-distribution of (a) maximum temperature  
and (b) minimum temperature

|  |  |
| --- | --- |
| (a) | (b) |

Figure 5 – LogGamma distribution of (a) maximum humidity temperature  
and (b) Beta distribution of minimum humidity temperature

# 6. Validation of Empirical and Theoretical Distributions Using Quantile Biplots

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like Student’s t-, Log-Gamma or Beta distribution.

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

|  |  |
| --- | --- |
| (a) Maximum temperature | (b) Minimum temperature |
| (c) Maximum relative humidity temperature | (d) Minimum humidity temperature |
| Figure 6 – QQ-plots | |

As we can see in Figure 5, all the points plotted on the graph of both maximum and minimum temperatures almost perfectly lie on a straight line, so we can clearly say that this is Student’s t-distribution. Similarly, we can talk about the Kurtosis (a measure of “Tailedness”) of the distribution by simply looking at its Q-Q plot. The distribution with a thin tail will form a Q-Q plot with a very less or negligible deviation at the ends.

As for humidity temperature distribution, it can be seen that the bottom end of the (c) QQ-plot deviates from the straight line but the upper end is not, so it can be clearly said that the distribution has a longer tail to its left or simply it is left-skewed (or *negatively skewed*).

# 7. Statistical Tests

|  |  |  |
| --- | --- | --- |
| **Target Variable’s Distribution** | **Kolmogorov-Smirnov test** | **Cramer–Von Mises**  **test** |
| Student’s t-distribution of maximum temperature |  |  |
| Student’s t-distribution of minimum temperature |  |  |
| Log-Gamma distribution of maximum humidity temperature |  |  |
| Beta distribution of maximum humidity temperature |  |  |

From the output of Kolmogorov-Smirnov for Student’s t-distribution we can see that the -value is . Since the -value is less than , the null hypothesis is rejected. It can be said that the sample data does *not* come from the Student’s t-distribution. The same can be concluded about others. Next, the chi-square test was performed to check whether there are differences between the observed (experimental) value and the expected (theoretical) value. As for all cases the -value is less than our chosen significance level , the null hypothesis is also rejected. It can be said that the sample data is different from theoretical one.

# 8. Sourcecode

Source code can be found in GitHub repository: github.com/ellkrauze/methodsmodels2022.