Analytics for Observational Data (IT142IU)

Lab 7: Bayesian statistics

## Objectives

* Understanding Bayes’ theorem, Bayesian inference
* Applying Bayesian inference to the existing datasets.
* Dataset sources:
  + <https://www.kaggle.com/datasets/fedesoriano/wind-speed-prediction-dataset>
  + <https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data>
* Programming languages: Python/Java
* Ref: Lecture notes in Session 10

## Tasks

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset | Wind\_dataset.csv |
| Reuse the random variable chosen in the previous lab. | Choosing WIND |
| Choose a good sample from the previous lab |  |
| Calculate Mean, variance, and number of the items in the sample data |  |
| Take TWO items and give their prior distributions for the mean value.  E.g. *p*1(*μ*) ~ *N*(11, 25) for the Wind | Create mean value prior distributions based on past information.  p1(u) ~ N (9.5, 26) p2(u) ~ N (10, 26.5) |
| Construct the posterior of the two cases above |  |
| Visualize the distributions of the two cases above |  |
| Remark | 1. Prior Beliefs:  - We began with assumptions on the mean wind speed of two objects:  - Item 1: We estimated that the variance was 26 and the average speed to be approximately 9.5.  - Item 2: We estimated that the variance was 26.5 and the average to be roughly 10.  - These were our educated assumptions before examining the real facts.  2. Updated Beliefs:  - Once we saw the sample data, our opinions shifted:  - Item 1: Based on our current estimation, the sample's average speed is approximately 9.94, which is extremely similar to what we saw, and the sample's variation is 0.03. We have a good deal of confidence in this.  - Item 2: This is the same tale. We are fairly certain that the sample variance is 0.03, and the average appears to be approximately 9.94.  - Thus, our revised beliefs are derived from our initial hypotheses as well as our observations from the evidence.  3. Conclusion: This research demonstrates how we updated our assumptions on the mean wind speed for two items using Bayesian inference. We were able to estimate the average speed with more accuracy when we combined our initial assumptions with the actual data. We appear to be reasonably confident in these estimations based on the slight differences in our updated beliefs. Overall, by taking into account both our previous views and new evidence, Bayesian methods assist us in making better judgments. |

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset | GlobalLandTemperaturesByCountry.csv |
| Choose a random variable | AverageTemperature |
| Choose a good sample from the previous lab |  |
| Calculate Mean, variance, and number of the items in the sample data |  |
| Take TWO items and give their prior distributions for the mean value.  E.g. *p*1(*μ*) ~ *N*(18, 100) for the average temperature |  |
| Construct the posterior of the two cases above |  |
| Visualize the distributions of the two cases above |  |
| Remark | - Effect of the Priors: The resulting posterior distributions have relatively similar means and variances, even if the prior distributions had distinct means and variances. This is because there are a lot of observed data points (1000), which significantly affects the posterior and lessens the influence of the prior.  - Posterior Means: The posterior distribution is heavily influenced by the data, as seen by the two-posterior means (16.92 for Prior 1 and 16.92 for Prior 2), which are both extremely near to the observed mean (16.924).  - Posterior Variances: Due to the vast quantity of data available, the posterior distributions are substantially more concentrated around the mean, as seen by the two posterior variances (0.13 and 0.13), which are both significantly smaller than the variances of the priors 130. |