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 | A portion of sample data with size of 1000 observations:  0 10.17  1 2.33  2 7.92  3 8.96  4 3.54 |
| Calculate Mean, variance, and number of the items in the sample data | 9.9023  24.8509  *n* = 1000 |
| Take TWO items and give their prior distributions for the mean value.  E.g. *p*1(*μ*) ~ *N*(11, 25) for the Wind | Construct prior distributions for mean value base on prior knowledge:  p1(u) ~ N (10, 25)  p2(u) ~ N (15, 30) |
| Construct the posterior of the two cases above | Construct posterior:  p1(u|x), ~ N (9.902347070436536, 0.024826198602184397)  p2(u|x), ~ N (9.906469290123708, 0.024830308216605323 |
| Visualize the distributions of the two cases above | Distribution of Prior 1:    Distribution of Prior 2: |
| Remark | * **Effect of Priors:** Despite the prior distributions being quite different in terms of their means and variances, the resulting posterior distributions are very similar, both in terms of their means and variances. This convergence is due to the large number of observations (1000 data points), which heavily influences the posterior, reducing the impact of the prior. * **Posterior Means:** Both posterior means (9.9023 for Prior 1 and 9.9065 for Prior 2) are very close to the observed mean (9.90225), indicating that the data has a strong influence on the posterior distribution. * **Posterior Variances:** Both posterior variances (0.0248) are significantly smaller than the variances of the priors (25 and 30), showing that the posterior distributions are much more concentrated around the mean, reflecting the large amount of data available. |

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset | GlobalLandTemperaturesByCountry.csv |
| Choose a random variable | AverageTemperature |
| Choose a good sample from the previous lab | A portion of sample data with size of 1000 observations:  0 0.460  1 -13.475  2 24.879  3 9.734  4 18.052 |
| Calculate Mean, variance, and number of the items in the sample data | 17.28348  120.11423  *n* = 1000 |
| 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 prior distributions for mean value base on prior knowledge:  p1(u) ~ N (18, 100)  p2(u) ~ N (20, 120) |
| Construct the posterior of the two cases above | Construct posterior:  p1(u) ~ N (18, 100)  p2(u) ~ N (20, 120) |
| Visualize the distributions of the two cases above | Distribution of Prior 1:    Distribution of Prior 2: |
| Remark | * **Effect of the Priors:** Despite the prior distributions having different means and variances, the resulting posterior distributions are very similar in terms of their means and variances. This is due to the large amount of observed data (1000 data points), which heavily influences the posterior, reducing the impact of the prior. * **Posterior Means:** Both posterior means (17.2843 for Prior 1 and 17.2862 for Prior 2) are very close to the observed mean (17.28348), indicating that the data strongly influences the posterior distribution. * **Posterior Variances:** Both posterior variances (0.11997 and 0.11999) are significantly smaller than the variances of the priors (100 and 120), showing that the posterior distributions are much more concentrated around the mean, reflecting the large amount of data available. |