
Problem assignment 2

Due: Thursday, February 3, 2022

Problem 1. Mean estimates and the effect of the sample size

In this problem we study the influence of the sample size on the estimate of the mean. The data for this experiment are in file *mean_study_data.txt* in the zip file linked to the assignment. The data were generated from the normal distribution with mean=15 and standard deviation=5. You were also given a python file hw-1-problem-1.py that loads the data from the text file and defines a function that lets you calculate the basic sample statistics such as mean, standard error and the confidence interval for a sample. You will need to extend this code to answer the questions bellow. Please submit the new version of the file with the code answering all the questions.

- (Part 1) Load the data in the *mean_study_data.txt*. Calculate and report the mean and standard deviation of the data. Compare them to the true mean and std above.
- (Part 2) Write a function subsample(data, k) that randomly selects k instances from the data in the *mean_study_data.txt*
- (Part 3) Use the above function to randomly generate 1000 subsamples of the data of size 25. For each subsample calculate its mean and save the results in the vector of 1000 means. Plot a histogram of 1000 mean values using 20 bins.
- (Part 4) Include the histogram in your report. Analyze the means calculated on 1000 subsamples of size 25 and compare them to the true mean and the mean that was calculated in step 1 on all examples in the dataset. Report your observations.
- (Part 5) Repeat steps from part 3 but now generate 1000 subsamples of size 40. Include the histogram in the report and compare it to the histogram generated in part 4 for subsamples of size 25, and to the mean of the original data. What are the differences? What conclusions can you make by comparing the means for subsamples of size 25 and 40.
- (Part 6) Take first 25 examples from the original data in the *mean_study_data.txt*. Use the function given to you that calculates the sample, sample standard error and its confidence interval. Use it to calculate the all stats for the first 25 examples and 0.95

confidence interval. Report the results. Does the true mean value (see first paragraph of the problem text) fall into the 0.95 confidence interval?

Problem 2. Probabilities

Part a. Assume you have 2 fair dice. What are the probabilities associated with the different outcomes that are obtained by summing together the numbers on the two dice?

Part b. Calculate the expected value of the outcome for the 2 fair dice roll experiment.

Part c. Assume you play the two dice game from part a. 5 times. What is the probability, we never see the outcome of 4? What is the probability we see odd-sum outcomes in all 5 trials.

Problem 3. Uniform distribution

Assume a uniform distribution $p(x|a,b) = \frac{1}{b-a}$ where $x \in [a,b]$.

- (a) Show that the distribution is properly normalized (that is, integral over its possible values equals 1)
- (b) Derive the mean of the distribution.

Problem 4. Bernoulli trials

Assume we have conducted a coin toss experiment with 100 coin flips. The results of the experiment are in file 'coin.txt' where 1 means a head and 0 means a tail. Assume that θ represents the probability of observing a head. Write and submit the python file hw-2-problem-4.py that loads the data and answers the following questions. Please report the answers also in your report.

- (a) What is the ML estimate of θ?
- (b) Assume the prior on θ is defined by a Beta distribution $Beta(\theta|1,1)$. Plot and report both the prior and the posterior distributions on θ. Please use scipy.stats libraries for the definition of Beta distribution.
- (c) Calculate and report:
  - the MAP estimate of θ based on the posterior from part b.
  - Expected value of θ based on the posterior from part b.

Show (plot) both the MAP estimate, and the expected value of θ on the plot of the posterior of θ you have generated in part b.

- (d) Repeat parts b and c by assuming that the prior on θ follows $Beta(\theta|4, 2)$.

Problem 5. Univariate Gaussian

Assume real valued measurements in file 'gaussian1.txt'. Write and submit a python code hw-2-problem-5.py that loads the data in the gaussian1.txt file and answers the following questions. Please use scipy.stats libraries for the Gaussian distribution models, and matplotlib for ploting the results.

- (a) Plot the histogram of the data using 10 bins.
- (b) Calculate and report the ML estimate of the mean and the variance from the data. Please use the unbiased estimate of the variance. Report the resulting Gaussian distribution: its mean and variance and plot the distribution.

Problem 6. Poisson distribution

The Poisson distribution is used to model the number of random arrivals to a system over a fixed period of time. Examples of systems in which events are determined by random arrivals are: arrivals of customers requesting the service, occurrence of natural disasters, such as floods, etc. The Poisson distribution is defined as:

$$p(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ for } x = 0, 1, 2, \cdots$$

where λ is a parameter.

Part 1. Answer the following question:

- (a) Assume we have n independent samples of x. What is the ML estimate of the parameter λ.

Part 2. Now we are ready to do some python experiments and plotting. Write python code that answers the questions below and include it in file hw-2-problem-6.py. Please use scipy.stats libraries for the poisson and gamma distribution models, and matplotlib for ploting the results.

Answer the following questions:

- (a) Plot and report the probability function for Poisson distributions with parameters $\lambda = 2$ and $\lambda = 6$. Note that the Poisson model is defined over non-negative integers only. The poisson model can be found in sc item (b) Assume the data in 'poisson.txt' that represent the number of incoming phone calls received over a fixed period of time. Compute and report the ML estimate of the parameter λ. Also plot and report the probability function for the ML parameter.

- (c) The conjugate prior for λ defining the Poisson distribution is Gamma distribution. It is defined as:

$$p(\lambda|a,b) = \frac{1}{b^a\Gamma(a)}\lambda^{(a-1)}e^{-\frac{\lambda}{b}}.$$

  Plot and report the Gamma distribution for the following set of parameters ($a = 1, b = 2$) and ($a = 3, b = 5$).

- (d) Assuming the prior distribution on λ is $Gamma(\lambda|a,b)$, the posterior distribution for λ after seeing observations $D = \{x_1, x_2, \ldots, x_n\}$ is again gamma distribution:

$$p(\lambda|D) \sim Gamma(\lambda|a + \sum_{i=1}^{n} x_i, \frac{b}{nb+1}).$$

  Please use data in 'poisson.txt' to calculate and plot the posterior distributions of λ for both priors in Part c.