Mouhammadou Dabo

February 11, 2021

CS 1675: Intro to Machine Learning

Professor Milos Hauskrecht

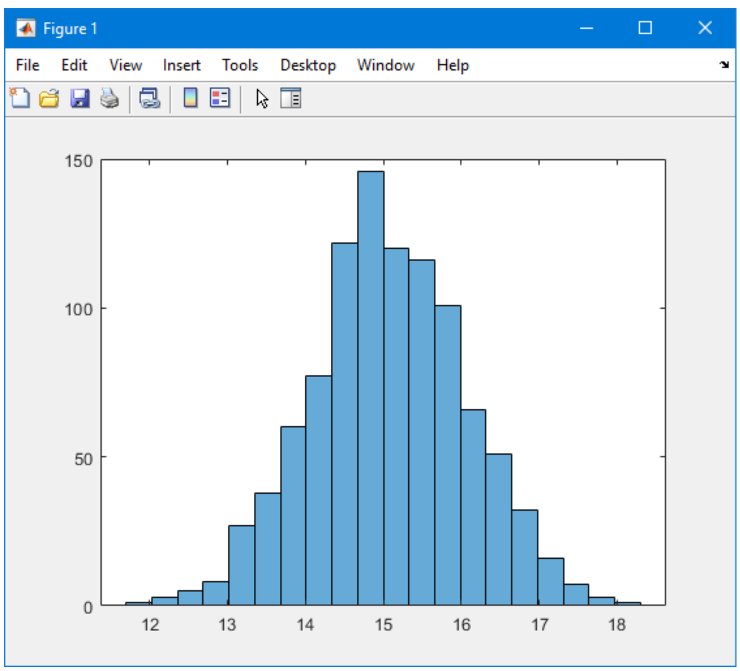
Handout 2 – Problem Assignment

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

Part 1

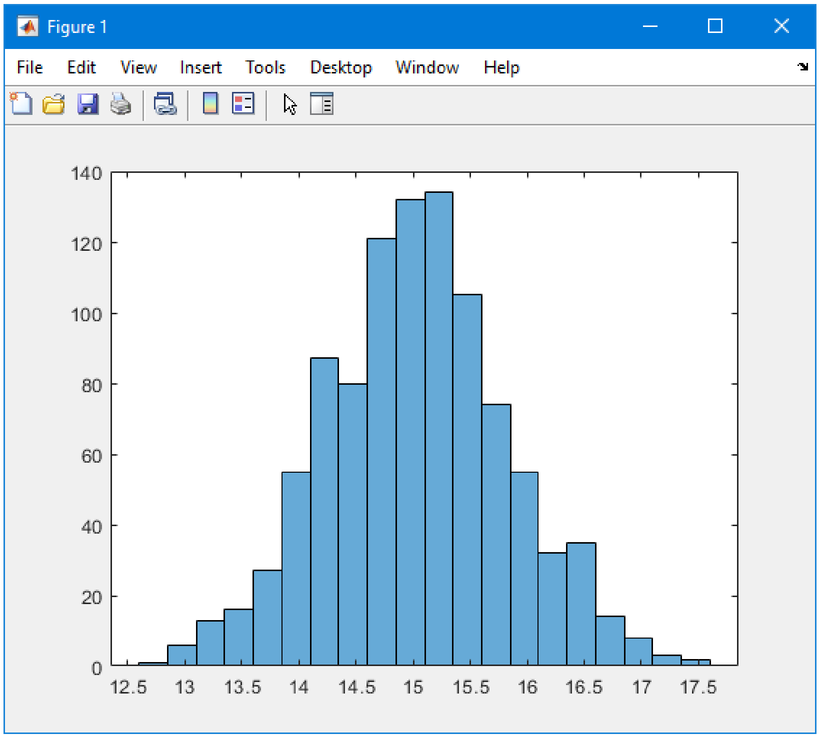
- The mean from *mean_study_data.txt* is 15.0415, while the standard deviation is 5.0279.
- Compared to the true mean and true standard deviation, this mean is 0.0415 higher and the standard deviation is 0.0279 higher.

Part 4



- The mean of the 1000 subsamples of size 25 was reported to be 15.0310.
- This new mean is 0.0105 less than the one reported in *mean_study_data.txt*, and 0.0310 higher than the true mean, making it closer to the true mean.

Part 5



- The histogram of this subsample compared to the one in part 4 has a slightly smaller range, and represents more of a normal distribution.
- The mean of this subsample is 15.0320, which is slightly higher than the mean of the subsamples of size 25.

Part 6

- After using the function t-test to calculate the confidence interval, the true mean does not fall into the 0.95 confidence interval.

Problem 2. k-fold cross-validation

Part 2

- Test 1
  - Mean: 3.993765
  - STD: 4.439093
- Test 2
  - Mean: 1.827680
  - STD: 3.627165
- Test 3
  - Mean: 2.144625
  - STD: 2.350383

- Test 4
    - Mean: 1.795354
    - STD: 3.159804

- Test 5
    - Mean: 2.084856
    - STD: 3.379448
- Test 6
    - Mean: 1.762692
    - STD: 3.264037
- Test 7
    - Mean: 2.104563
    - STD: 3.462198
- Test 8
    - Mean: 1.034257
    - STD: 2.580015
- Test 9
    - Mean: 1.583655
    - STD: 3.418371
- Test 10
    - Mean: 2.424611
    - STD: 2.283125

Problem 3. Probabilities

Part a.

- The possible outcomes of rolling the 2 fair dice are:
    - (1,1) (1,2) (1,3) (1,4) (1,5) (1,6)
    - (2,1) (2,2) (2,3) (2,4) (2,5) (2,6)
    - (3,1) (3,2) (3,3) (3,4) (3,5) (3,6)
    - (4,1) (4,2) (4,3) (4,4) (4,5) (4,6)
    - (5,1) (5,2) (5,3) (5,4) (5,5) (5,6)
    - (6,1) (6,2) (6,3) (6,4) (6,5) (6,6)
- With the possible outcomes of summing the two dice and their probabilities being:
    - 2 with a probability of 1/36
    - 3 with a probability of 2/36 = 1/18
    - 4 with a probability of 3/36 = 1/12
    - 5 with a probability of 4/36 = 1/9
    - 6 with a probability of 5/36
    - 7 with a probability of 6/36 = 1/6
    - 8 with a probability of 5/36

- 9 with a probability of 4/36 = 1/9
- 10 with a probability of 3/36 = 1/12
- 11 with a probability of 2/36 = 1/18
- 12 with a probability of 1/36

Part b.

- The expected value of the outcome for the 2 fair dice roll experiment can be found by adding all the possible outcomes and dividing by the possible combinations:
  - $(2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 11 + 12) \times \frac{1}{36}$
  - $= 252 \times \frac{1}{36}$
  - $= 7$
- Therefore, the expected value is 7.

Part c.

- The probability of never seeing the outcome of 4 can be found by subtracting the total probability of 1 with the probability of seeing the outcome 4, and then multiplying itself 5 times:
  - $\frac{36}{36} - \frac{4}{36} = \frac{32}{36} = \frac{8}{9}$
  - $\frac{8}{9} \times \frac{8}{9} \times \frac{8}{9} \times \frac{8}{9} \times \frac{8}{9} = \frac{32{,}768}{59{,}049}$
  - Therefore, the probability of never seeing the outcome of 4 after playing 5 times is 32,768/59,049 which is about 0.55493.
- The probability of seeing an odd-sum outcome can be found by adding all the odd-sum probabilities and then multiplying that probability by itself 5 times:
  - $\frac{2}{36} + \frac{4}{36} + \frac{6}{36} + \frac{4}{36} + \frac{2}{36} = \frac{18}{36} = \frac{1}{2}$
  - $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{32}$
  - Therefore, the probability of seeing an odd-sum outcome in all 5 trials would be 1/32 which is about 0.03125.

Problem 4. Probabilities: Bayes theorem

- *P(disease = T)* = 0.0001 (incidence of disease on population)
- *P(disease = F)* = 0.9999 (non-incidence of disease on population)
- *P(Negative | T)* = 0.01 (test is negative, but disease is still present)
- *P(Positive | T)* = 0.99 (test is positive, and disease is present)
- *P(Negative | F)* = 0.99 (test is negative, and disease is not present)
- *P(Positive | F)* = 0.01(test is positive, but disease is not present)

P(T | Positive) = $\frac{P(Positive \mid T) \times P(T)}{P(Positive)}$

where

- *P(Positive) = P(Positive | T) x P(T) + P(Positive | F) x P(F)*
- *= 0.99 x 0.0001 + 0.01 x 0.9999 = 0.010098*

So

- *P(T | Positive)* = $\frac{0.99\ x\ 0.0001}{0.010098} = 0.0098$

According to the above probability of somebody from the wide population testing positive and indeed suffering from the disease is very, very low, so I would not recommend the test the whole population.

Problem 5. Uniform distribution

(a)

Given the distribution, when $x \in [a, b]$

$$f(x) = \frac{1}{b-a}, where\ a \le x \le b$$
$$F_x(x) = \int_{-\infty}^{x} f(x)dx$$
$$= \int_a^x \frac{1}{b-a} du = \frac{u}{b-a} |_a^x$$
$$= \frac{x}{b-a} - \frac{a}{b-a} = \frac{x-a}{b-a}$$

So, the CDF at x = b is

$$,F_x(b) = \frac{b-a}{b-a} = 1$$

(b)

To derive the mean of the distribution,

$$E(x) = \int_{-\infty}^{\infty} x * f(x)dx$$
$$= \int_{-\infty}^{\infty} x * \left(\frac{1}{b-a}\right) dx$$
$$= \left(\frac{1}{b-a}\right) \int_a^b x * dx = \left(\frac{1}{b-a}\right) \left[\frac{x^2}{2}\right] |_a^b$$
$$= \left(\frac{1}{b-a}\right) \left(\frac{b^2}{2} - \frac{a^2}{2}\right) = \left(\frac{1}{b-a}\right) \frac{(b+a)(b-a)}{2}$$
$$= \frac{b+a}{2}$$

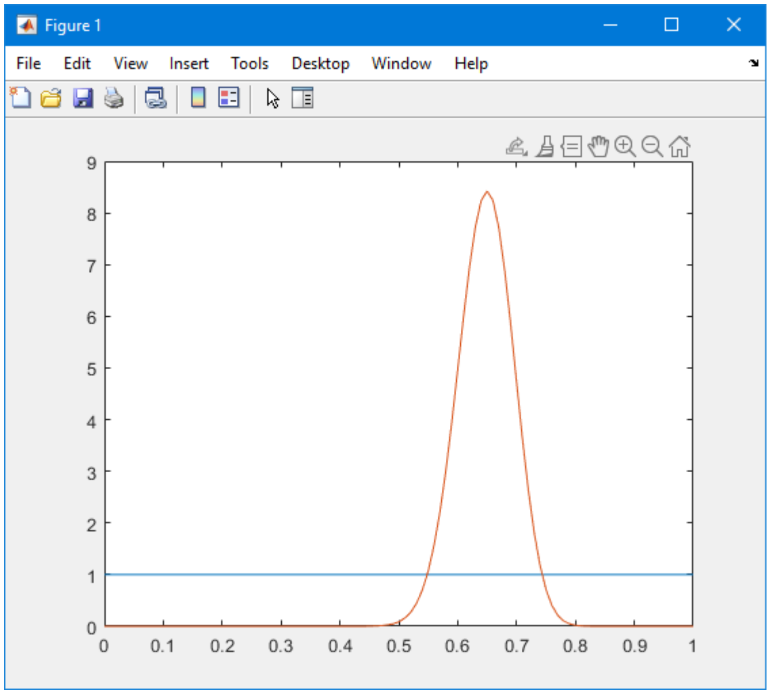
Making the mean of the distribution equal to (b + a) / 2.

Problem 6. Bernoulli trials

(a)

The ML estimate of theta is 0.65

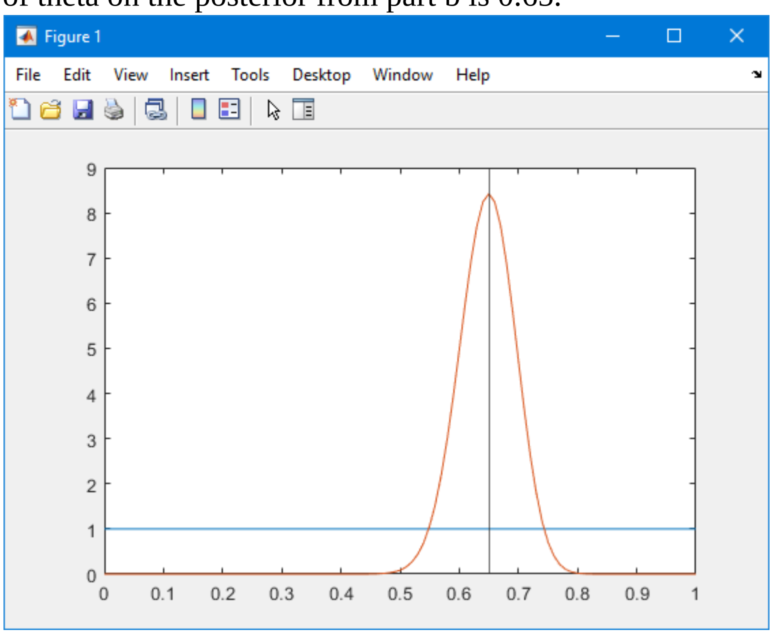
(b)



The prior is represented by the blue equation, while the posterior is represented by the red equation.

(c)

MAP estimate of theta on the posterior from part b is 0.65.



(d)

MAP estimate of theta on the posterior from part b is 0.6538.