EE 503 : Homework 6

Due : 10/31/2023, Tuesday before class.

1. An important manufacturing process produces cylindrical component parts for the automotive industry. It is important that the process produce parts having a mean diameter of 5 mm. The engineer involved conjectures that the population mean is 5.0 mm. An experiment is conducted in which 100 parts produced by the process are selected randomly and the diameter measured on each. It is known that the population standard deviation $\sigma = 0.1$. The experiment indicates a sample average diameter $\overline{x} = 5.027$ mm. Does this sample information appear to support or refute the engineer's conjecture?

2. A random variable X is believed to be uniformly distributed in the interval $[0, a]$, where a is unknown. Derive the maximum likelihood estimate, $\hat{a}$, of a from n independent observations of X labelled X_1, ..., X_n. Find the mean and variance of $\hat{a}$ and explain why it is a good estimate of a when n is large.

3. A communication channel accepts as input either 000 or 111. The channel transmits each bit (0 or 1) correctly with probability $1 - p$ and erroneously with probability p. Find the entropy of the input given that the output is 010.

4. A transmitter generates a set of messages $m \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ with probabilities $\{1/4, 1/4, 1/8, 1/8, 1/8, 1/16, 1/32, 1/32\}$, respectively.

   a) What is the entropy of m?

   b) For binary transmission, n distinct messages can be represented using $\lceil \log_2 n \rceil$ bits. E.g., the 8 messages in m can be represented uniquely using 3 bits with the following mapping $\{1, 2, 3, 4, 5, 6, 7, 8\} \rightarrow \{001, 010, 011, 100, 101, 110, 111, 000\}$. Such a mapping is called a *code* and each 3-bit representation of a message is called a *codeword.* For the above code, the average length of a codeword is 3 bits since each codeword is 3 bits long. Can you design another code for m such that the average length of a codeword is equal to the entropy of m? (*Note:* $\lceil . \rceil$ denotes the ceiling function.)

5. **Simulation Problem**: In statistical inference, one wishes to estimate unknown population parameters θ (for example, the population mean) using observed sample data. A *confidence interval* is a random interval calculated from the sample data that contains θ with a specified probability. The interval has an associated *confidence level* C which gives the probability that any valid confidence interval will contain the true value of the unknown parameter θ. For example, assume that the confidence level is 95%. Then, if we construct 100 different 95% confidence intervals for θ using 100 independent sets of data, then we would expect about 95 of them to contain θ.

a) Use *random.expovariate(*λ*)* from Python's random library to create $n = 5000$ samples of an exponential distribution with mean 3. Note that in a real-world scenario, the true distribution of the sample data is not known.

b) Compute the sample mean $\bar{x}$ of the data.

c) Assume that the true standard deviation σ of the exponential distribution is given to you. Use a confidence level of $C = 95\%$ to compute $z^* = \Phi^{-1}(1 - \frac{\alpha}{2})$, where $\alpha = 1 - \frac{C}{100}$, and Φ is the CDF of the standard normal distribution (see the figure below). Build the confidence interval $I = (\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}})$. Does this interval contain the true mean of the exponential distribution? You can use existing tables in order to compute Φ^{-1}.

d) Repeat parts a - c for 10 different sets of $n = 5000$ samples. How many of the confidence intervals contain the true mean of the exponential distribution?

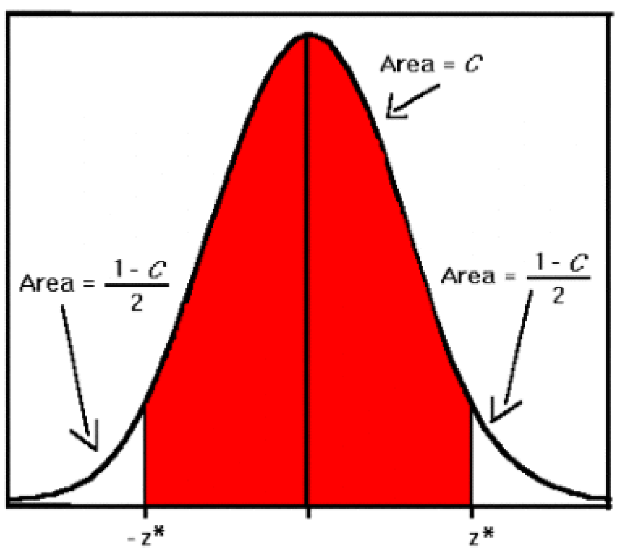

Figure 1: Interpretation of confidence level