ECE 131A
Probability and Statistics
Instructor: Lara Dolecek
TAs: Lev Tauz
Debarnab Mitra

Class Project

Due: March 14, 2021
levtauz@ucla.edu
debarnabucla@ucla.edu

Please upload your project to Gradescope by March 14, 11:59 pm PST.
Please submit a single PDF directly on Gradescope
You may type your project report or scan your handwritten version. Make sure all the work is discernible.
All MATLAB code shall be submitted in a single zip file on CCLE.
100 points total

In this project we will further analyze random variables and learn about their utility in practical systems. Each part will have a combination of programming, mathematical analysis, and technical writing. You will be graded on all components.

When producing your plots **clearly indicate** the x-axis, the y-axis and what is being plotted (using legends, title etc.). You may need to rescale x-axis to ensure that your plot is showing the right quantity. All plots are to be generated in MATLAB.

The required programming language for this project is **MATLAB**, you **may not use any other programming language**. Make sure to attach in the appendix of your project report **all programs/code** that you used to generate the data and plots.

There is no collaboration for the project. You may discuss with other students about the project but all writing and code must be **your own work**. Project report and code will be checked for plagiarism.

1. (25 pts) *Tossing a fair and unfair die.* Suppose you have a 5-sided die, with sides numbered 1, 2, 3, 4, and 5.

   (a) Write a MATLAB program to simulate the tossing of a 5-sided fair die, for t =10, 50, 100, 500 and 1000 tosses. Based on the simulation, what is the estimated probability of obtaining an odd number?

   Solution:

   ```
   %Estimated Probability of obtaining an odd number with N=50 is 0.68
   %Estimated Probability of obtaining an odd number with N=100 is 0.56
   %Estimated Probability of obtaining an odd number with N=500 is 0.588
   %Estimated Probability of obtaining an odd number with N=1000 is 0.599
   ```

   (b) Suppose X is a random variable denoting the outcome of a die toss. Based on the mathematical analysis, what is the probability that X has odd value?
   Solution: Theoretically, the probability that X takes odd value is 0.6.

   (c) Refer back to part (a). Does it agree with the theoretical result in (b)?

   Solution: The empirical estimation match the theoretical result. The estimate is more precise as N gets larger.

   (d) Repeat parts (a), (b), and (c) with a 5-sided die that has the following properties:

   - The probability of the die outcome being 1 is equal to the probability of the die outcome being 2.
   - The probabilities of the die outcome being a 3, 4, or 5 are all equal.
   - The probability of the die outcome being a 1 is twice the probability of the die outcome being 5.

   Solution:

   ```
   %Estimated Probability of obtaining an odd number with N=50 is 0.52
   %Estimated Probability of obtaining an odd number with N=100 is 0.54
   %Estimated Probability of obtaining an odd number with N=500 is 0.566
   %Estimated Probability of obtaining an odd number with N=1000 is 0.583
   ```

   Theoretically, the probability that X takes odd value is 0.5714. The empirical estimation match the theoretical result. The estimate is more precise as N gets larger.

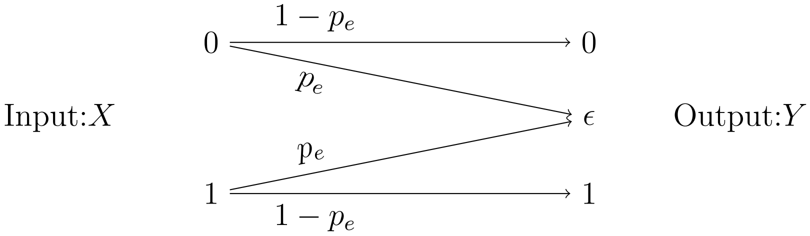
   You may find useful the MATLAB function `rand` that generates a uniform random value in the $(0, 1)$ interval.

2. (25 pts) *Coding for BEC*

In this problem, we consider the problem of transmitting bits over a binary erasure channel (BEC). The input bit X to the BEC channel is "0" or "1" with equal probability. The output of the channel Y is given by

$$Y = \begin{cases} X & \text{wp } 1 - p_e \\ \epsilon & \text{wp } p_e \end{cases}$$

where p_e is called the erasure probability and the bits that become ϵ are said to be erased. Note that an observer of Y will know when a bit has been erased. We can summarize the communication channel as follows:



Clearly, if we send just one bit of information, the probability that the Output gets the Input is $1-p_e$. To improve the probability of successfully sending information through the BEC channel, we employ the ideas of redundancy and coding which turn out to be very powerful tools in solving such problems.

Consider the following strategy of using the N*-repetition code* described in Lecture 5. The N*-repetition code* takes a single bit that we want to transmit and repeats it N times (known as encoding). Now, each bit of the encoded N-bit sequence is transmitted through the BEC channel. Upon receiving this sequence, the receiver declares that the encoded bit is i) 1 if the received bit sequence has at least one 1, ii) 0 if the received bit sequence has at least one 0. If the received sequence has all erasures, the receiver declares that the encoded bit is undecodable. This process is known as decoding.

Consider the following example. Let $N = 5$ and suppose we want to transmit the bit 1. We first encode it into the sequence 11111, using the 5-repetition code. Suppose some bits got erased during transmission over the BEC channel, so that the receiver observes the sequence $\epsilon\epsilon\epsilon\epsilon 1$. In this example, using the decoding method, the receiver would determine that the encoded bit for transmission was a 1. If the receiver observes the sequence $\epsilon\epsilon\epsilon\epsilon\epsilon$,the receiver declares the encoded bit as undecodable.

(a) Write a program to simulate the BEC channel. Also, write programs for encoding and decoding the N-repetition code. We choose $N = 3, 4$ and 5 for this problem. For a given N, simulate 100000 instances of sending a bit through the BEC using the repetition code for $p_e = \{0.125, 0.15, 0.175, \cdots, 0.4\}$. Record the fraction of times (of the 100000), the receiver declares the encoded bit as undecodable for each p_e. Denote the fraction by P_{sim}^{Rep}. Plot P_{sim}^{Rep} against p_e for each N. You may want to use a semi-log for the y axis.

Solutions:

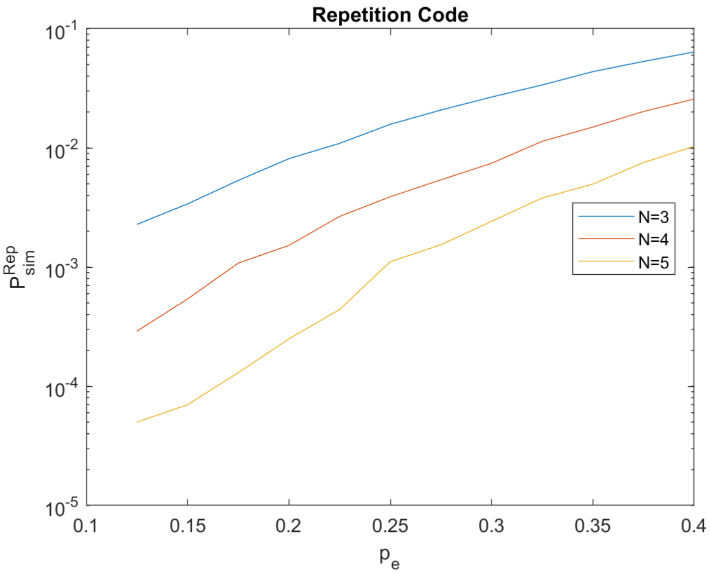


Figure 1: P_{sim}^{Rep} vs p_e

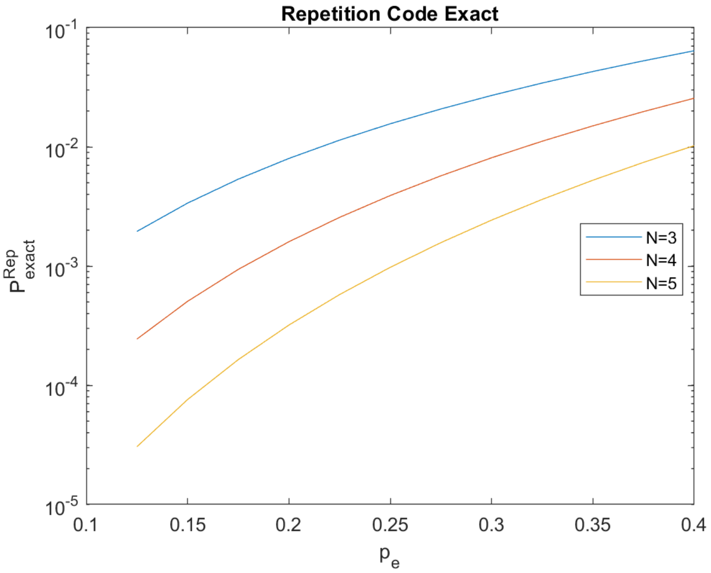


Figure 2: P_{exact}^{Rep} vs p_e

(b) Find the theoretical probability of the N-repetition code being undecodable and denote it as P_{exact}^{Rep}. Your answer should be parameterized by N and p_e. For $N = 3, 4$ and 5 and $p_e = \{0.125, 0.15, 0.175, \cdots, 0.4\}$, generate plots similar to (a) using your theoretical results for P_{exact}^{Rep} against p_e for each N.

Solutions:

We can consider the transmitting of a single bit as a Bernoulli trial with 1 indicating that the bit is erased and 0 meaning that it got through. Let Y be the number of erasures in the N-repetition code. Observe that Y is a binomial random variable with parameters $n = N$ and $p = p_e$. Clearly, the bit is undecodable when all N message bits are erased. As such, $P_{exact}^{Rep} = P(Y = N) = p_e^N$.

Now, we consider a new strategy known as the *N-Single-parity code*. The *N-Single-parity code* takes a bit sequence of length $N-1$, and appends to it the modulo 2 sum of the $N-1$ bits to get a N-bit sequence. Specifically, if $x_1x_2 \ldots x_{N-1}$ is the sequence we want to transmit, the encoded sequence will be $x_1x_2 \ldots x_{N-1}x_N$ where $x_N = \sum_{i=1}^{N-1} x_i \mod 2$. This is the encoding process for the *N- single parity code.* Now, each bit of the N-bit sequence is transmitted through the BEC channel. Observe that a Single-parity code can reconstruct any single erasure. Specifically, if we assume x_j is erased and $j \neq N$, then we can reconstruct x_j by $x_j = x_N + \sum_{i=1, i\neq j}^{N-1} x_i \mod 2 = \sum_{i=1}^{N-1} x_i + \sum_{i=1, i\neq j}^{N-1} x_i \mod 2$. If x_N was erased, we can reconstruct it by simply summing the first $N-1$ bits again. As such, we claim the receiver can only decode if there is at most one erased bit in the received sequence. Thus, we get the following decoding process: i) If there are more than one erased bits in the received sequence, the receiver declares the encoded bits (of $N-1$ length) as undecodable; ii) if there is a single erased bit, the receiver replaces the erased bit by the method previously described. The first $N-1$ bits of this sequence is declared as the encoded bits by the receiver.

Consider the following example. Let $N = 5$ and suppose we want to transmit 0011. Using the 5-single parity code, we first encode 0011 into the sequence 00110, since the modulo 2 sum of 0011 is 0. Suppose some bits got erased during transmission over this BEC channel, so that the receiver observes the sequence $0\epsilon110$. In this example, since there is only one erasure, the receiver would replace the ϵ by the modulo 2 sum of the remaining bits 0110, which is 0. Thus the receiver forms the 5 length sequence 00110, and declares the first 4 bits of the sequence, i.e, 0011 as the original encoded bits. If the receiver observes the sequence $0\epsilon\epsilon10$ the receiver declares the encoded bits as undecodable as there are more than one erased bits.

(c) Repeat parts (a) and (b) using the *Single-parity code.* Denote P_{sim}^{Sin} as the fraction of undecodable messages and P_{exact}^{Sin} as the theoretical probability that a Single-parity code is deemed undecodable.

Solutions:

Plots:

Similarly to part(b), we can consider the transmitting of a single bit as a Bernoulli trial with 1 indicating that the bit is erased and 0 meaning that it got through. Let Y be the number of erasures in the Single-parity code. Observe that Y is

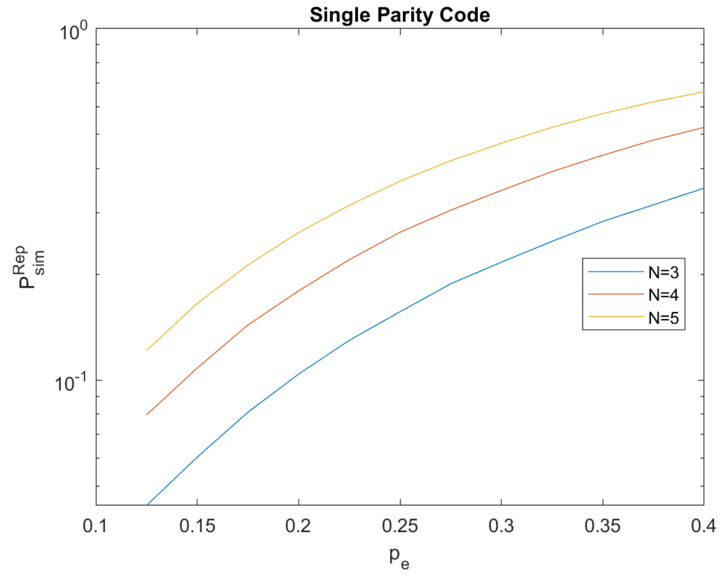


Figure 3: P_{sim}^{Rep} vs p_e

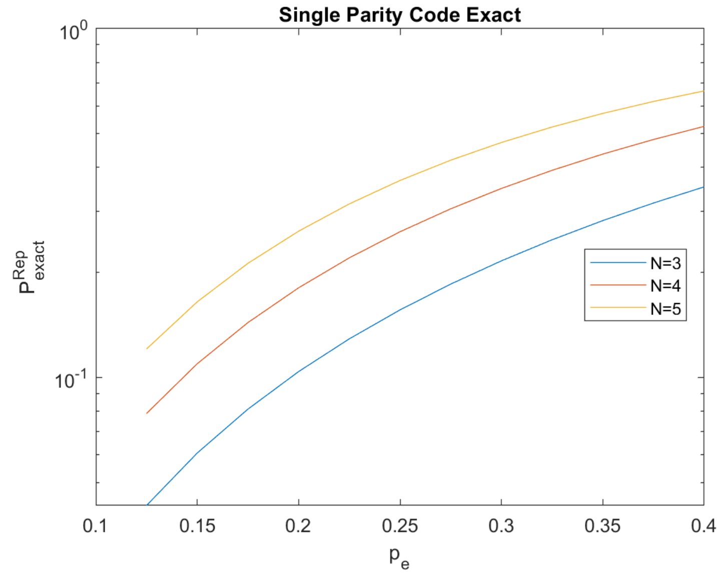


Figure 4: P_{exact}^{Rep} vs p_e

a binomial random variable with parameters $n = N$ and $p = p_e$. Clearly, the encoded bits are undecodable if there are at least two erasures. As such,

$$P_{exact}^{Sin} = P(Y \geq 2) = 1 - P(Y = 0) - P(Y = 1) = 1 - (1 - p_e)^N - Np_e(1 - p_e)^{N-1}.$$

(d) Compare the plots of P_{sim}^{Rep} and P_{sim}^{Sin}. Which code has a a higher probability of being undecodable in general?

Solutions: In general, the Single-parity code has a higher probability of being undecodable. Intuitvately, we can see that this is true because the N-repetition code only fails when all the bits have been erased which is also true for the Single-parity code. Thus, the Single-parity code must be worse in probability of being undecodable for all N.

(e) Under what scenarios would you use the Repetition Code over the Single-parity code and vice-versa?

Solutions:

From the previous part, we can see that we want to use the Repetition code over the Single-parity code whenever we prioritize the message getting through the channel since the Repetition code has a much lower probability of failing to decode.

On the other hand, the Repetition code only transmits 1 bit of useful information for every N bit that go through the channel. Conversely, the Single-parity code transmits $N-1$ useful bits for every N bits sent trough the channel. So we can see that when the Single-parity bit code succeeds, it has a much higher throughput of useful information than the Repetition code. Thus, one would use the Single-parity bit code when throughput is important.

3. (25 pts) *Naïve Bayes Classifier.* To maximize profit, advertisement companies like Google, Facebook, and Amazon want to target their advertisement to each user. Specifically, given a user and a product to sell, they want to determine whether this user will buy it. In this problem, we will build a classifier to help us determine whether a user will buy a certain product by using historical data.

   Some demographic data is collected for 887 users and are provided in `user_data.csv`. This dataset has four fields: i) Whether this user bought the product (1 for did buy and 0 for did not buy), ii) Type of Spender (1 for larger spender, 2 for medium spender, 3 for small spender) iii) Sex of user (1 for Male and 0 for Female), iv) Age of user. We use random variables B, T, S, and A for bought status, type of spender, sex, and age, respectively. In this problem, we are going to build a popular *Naïve Bayes Classifier* to predict B given T, S, and A.

   (a) Estimate the individual PMFs for B, T, S, and A by finding the fraction of each realization of these random variables among all data. Plot these PMFs (i.e. 4 PMFs in total).

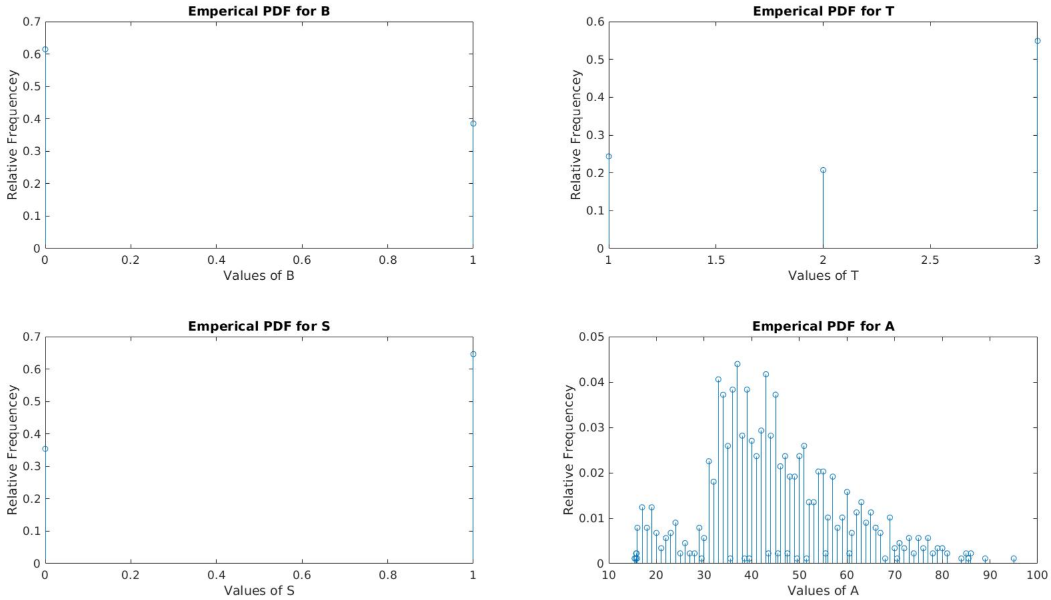
   (b) We are interested in how T, S, and A affect B, in the context of the *Naïve Bayes Classifier*; the first step is to estimate the conditional PMF conditioning on the outcome of interest, i.e., survival or not. Estimate and plot the conditional PMFs for T, S, and A separately conditioned on $B = 1$ and $B = 0$ (i.e. 6 PMFs in total).

   (c) In the *Naïve Bayes Classifier*, we use the conditional independence assumption. For example, if T, S, and A are conditionally independent on B, then $P(T, S, A|B = 0) = P(T|B = 0)P(S|B = 0)P(A|B = 0)$ and $P(T, S, A|B = 1) = P(T|B = 1)P(S|B = 1)P(A|B = 1)$. Using this assumption, compute $P(B = 0, T = 1, S = 0, A \leq 55)$ and $P(B = 1, T = 1, S = 0, A \leq 55)$ based on your estimations in (a) and (b).

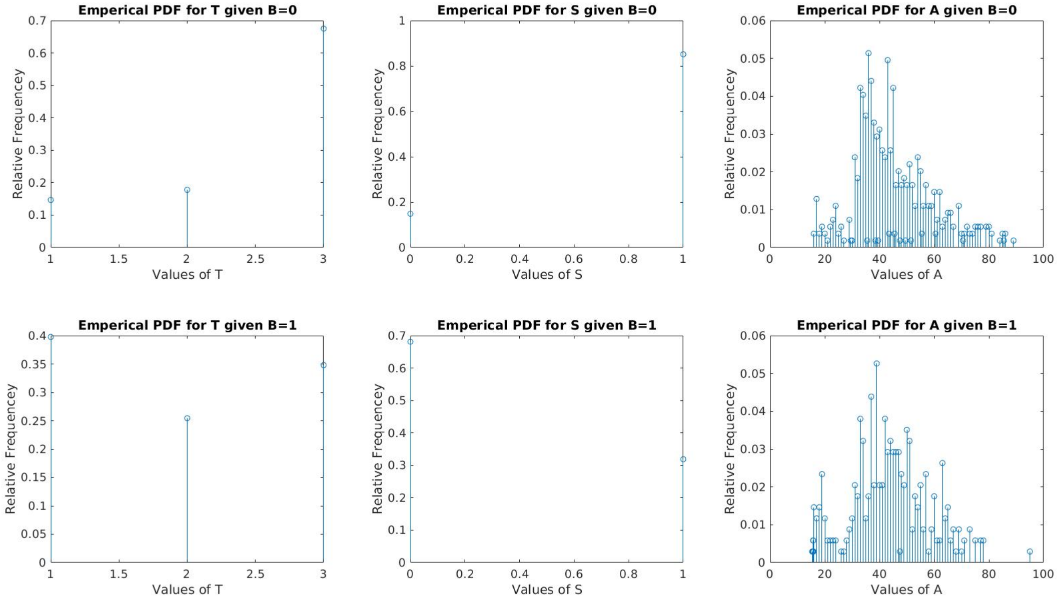
   (d) Based on your result in (c), compute $P(B = 0|T = 1, S = 0, A \leq 55)$ and $P(B = 1|T = 1, S = 0, A \leq 55)$. Predict whether a female whose age is below 55 and who is a large spender will buy this product or not.

   Solution:

   (a) Plot for part a:



(b) Plot for part b:



(c) We know the following from part (a) and (b):

$$\begin{aligned}
P(B=0) &= 0.6144;\\
P(T=1|B=0) &= 0.1468;\\
P(S=0|B=0) &= 0.1486;\\
P(A\leq 55|B=0) &= 0.7927;\\
P(B=1) &= 0.3856;\\
P(T=1|B=1) &= 0.3977;\\
P(S=0|B=1) &= 0.6813;\\
P(A\leq 55|B=1) &= 0.8099;
\end{aligned}$$

With these probabilities, we calculate $P(B=0, T=1, S=0, A\leq 55)$ and $P(B=1, T=1, S=0, A\leq 55)$ as follows:

$$\begin{aligned}
&P(B=0, T=1, S=0, A\leq 55)\\
=&P(B=0)\times P(T=1, S=0, A\leq 55|B=0)\\
=&P(B=0)\times P(T=1|B=0)\times P(S=0|B=0)\times P(A\leq 55|B=0)\\
=&0.0106.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&P(B=1, T=1, S=0, A\leq 55)\\
=&P(B=1)\times P(T=1, S=0, A\leq 55|B=0)\\
=&P(B=1)\times P(T=1|B=1)\times P(S=0|B=1)\times P(A\leq 55|B=1)\\
=&0.0846.
\end{aligned}$$

(d) From part (c), we already calculated the joint probabilities, therefore

$$\begin{aligned}
&P(B=0|T=1, S=0, A\leq 55)\\
=&\frac{P(B=0, T=1, S=0, A\leq 55)}{P(B=0, T=1, S=0, A\leq 55)+P(B=1, T=1, S=0, A\leq 55)}\\
=&\frac{0.0106}{0.0106+0.0846}=0.1113.
\end{aligned}$$

$$\begin{aligned}
&P(B=1|T=1, S=0, A\leq 55)\\
=&\frac{P(B=1, T=1, S=0, A\leq 55)}{P(B=0, T=1, S=0, A\leq 55)+P(B=1, T=1, S=0, A\leq 55)}\\
=&\frac{0.0846}{0.0106+0.0846}=0.8887.
\end{aligned}$$

We therefore predict that this women will buy the product.

4. (25 pts) *Central Limit Theorem.* Let X_1,X_2, X_3... be a sequence of i.i.d. random variables with finite mean μ and finite variance σ^2, and let Z_n be the mean of the first n random variables in the sequence:

$$Z_n = \frac{X_1 + X_2 + ... + X_n}{n}.$$

(a) Let X_i, for $i = 1, 2, \ldots$ be a uniform continuous random variable taking values in the interval $(3, 7)$. Write a MATLAB program to plot the pdf of Z_n. Consider $n = 1, 2, 3, 10, 30, 100$ and compare your results across different n's.

Solution As n increases, the pdf behave more like the gaussian.

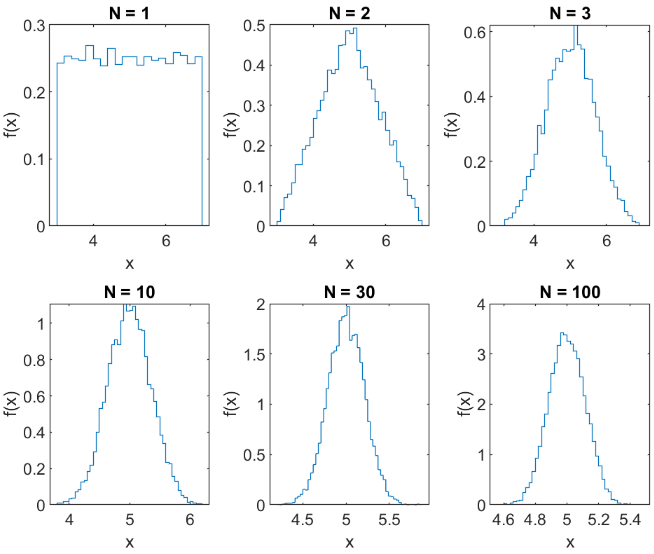


Figure 5: 4(a) PDF of Z_n for different n.

(b) Calculate analytically the mean and the variance of X_i and of Z_n in part (a).

Solution:

$E[X_i] = 5$, $Var(X_i) = 1.33$
$E[Z_n] = 5$, $Var(Z_n) = 1.33/n$

(c) Write a MATLAB program to generate a Gaussian random variable with the same mean and variance as Z_n. Superimpose its pdf on the plots from part (a).
Solution

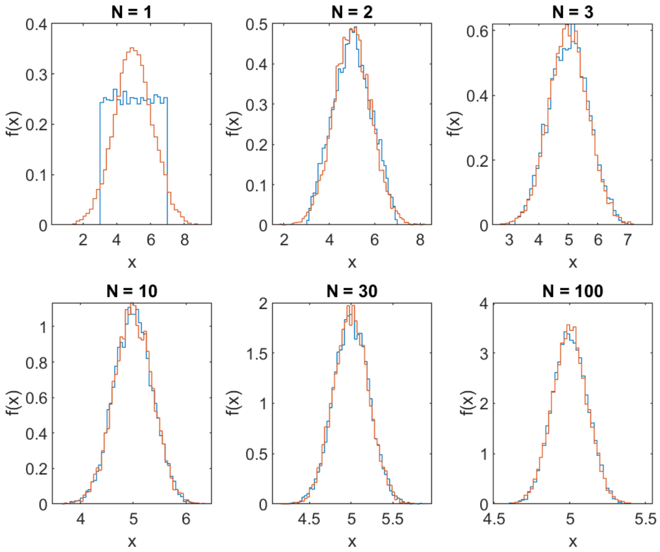


Figure 6: $4(c)$ PDF of Z_n and superimposed gaussian for different n.

(d) Repeat parts (a), (b), and (c) with X_i representing a toss of a 5-sided that is described in Problem 1(d). Note that X_i and Z_n are discrete in this case.

Hint. You can calculate the PDF analytically or empirically. For the later method, use $t = 10^4$ samples and while plotting histogram for discrete data, use `'BinWidth'` as $\frac{1}{n+1}$.

Solution

$E[X_i] = 2.57$, $Var(X_i) = 1.95$
$E[Z_n] = 2.57$, $Var(Z_n) = 1.95/n$

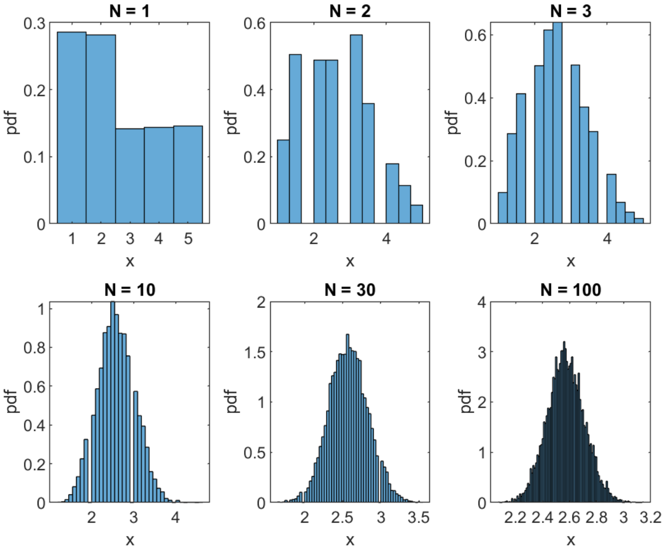


Figure 7: 4(d) PDF of Z_n for different n.

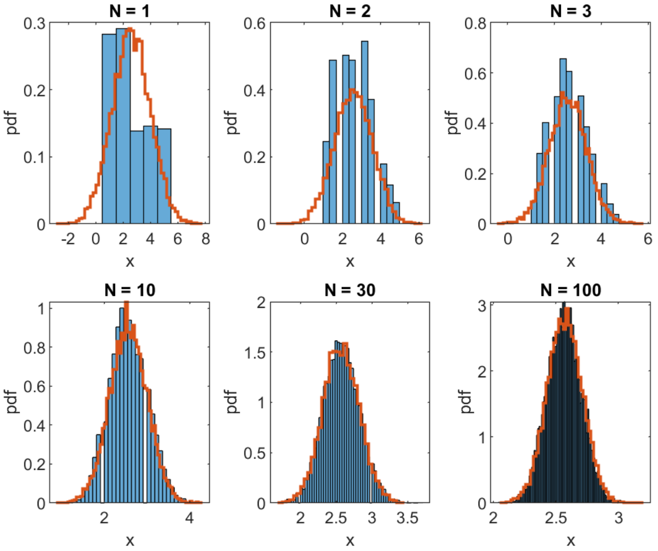


Figure 8: 4(d) PDF of Z_n and superimposed gaussian for different n.