1. *Program A will run 20 algorithms in sequence with the running time for each algorithm being independent random variables with a mean of 50 seconds and a variance of 100 seconds^2. Program B will run 20 algorithms in sequence with the running time for each algorithm being independent random variables with a mean of 52 seconds and variance of 200 seconds^2.*

   (a) *What is the approximate probability that Program A completes in less than 950 seconds?*

   Y_a is the total run time for Program A
   X_a is the run time for an algorithm in Program A

$$\begin{aligned} X_a &\sim N(\mu = 50 \text{ seconds}, \sigma^2 = 100 \text{ seconds}^2) \\ Y_a &\sim \sum_{i=1}^{20} X_{a,i} = 20\bar{X}_a \\ Y_a &\sim N(\mu = 1000 \text{ seconds}, \sigma^2 = 2000 \text{ seconds}^2) \\ P(Y_a < 950) &= \Phi(\frac{y_a - \mu_{Y_a}}{\sigma_{Y_a}}) = .1318 \end{aligned}$$

   (b) *What is the approximate probability that Program B completes in less thatn 950 seconds?*

   Y_b is the total run time for Program B
   X_b is the run time for an algorithm in Program B

$$\begin{aligned} X_b &\sim N(\mu = 52 \text{ seconds}, \sigma^2 = 200 \text{ seconds}^2) \\ Y_b &\sim \sum_{i=1}^{20} X_{a,i} = 20\bar{X}_b \\ Y_b &\sim N(\mu = 1040 \text{ seconds}, \sigma^2 = 4000 \text{ seconds}^2) \\ P(Y_b < 950) &= \Phi(\frac{y_b - \mu_{Y_b}}{\sigma_{Y_b}}) = .0774 \end{aligned}$$

   (c) *What is the approximate probability that Program A completes in less time than Program B?*

$$\begin{aligned} Y_b - Y_a &\sim N(\mu = \mu_{Y_b} - \mu_{Y_a}, \sigma^2 = \sigma_{Y_b}^2 + \sigma_{Y_a}^2) \\ P(Y_a - Y_b \leq 0) &= \Phi(\frac{(y_a - y_b) - \mu}{\sigma}) = .1855 \end{aligned}$$

2. *A fair 6-sided die is repeatedly rolled until the total sum of all the rolls exceeds 100. Approximate the probability that at least 30 rolls are necessary to reach a sum that exceeds 100.*

   X is single roll of the die
   Y is the sum of all rolls of the die

n is the number of rolls, which is equal to 30

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\bar{X} \sim N(\mu = E[X_i], \sigma^2 = \frac{Var(X_i)}{n}) \text{ as } n \to \infty \qquad \text{by Central Limit Theorem}$$

$$Y = \sum_{i=1}^{n} = n\bar{X}_i \sim N(\mu = nE[X_i], \sigma^2 = nVar(X_i))$$

$$E[X_i] = \sum_{i=1}^{6} \frac{i}{6} = \frac{7}{2}$$

$$Var(X_i) = E[X_i^2] - E[X_i]^2 = \frac{35}{12}$$

$$P(Y \leq 100) = \Phi(\frac{y - \mu_Y}{\sigma_Y}) = .2965$$

3. *From past experience, we know that the midterm score for a student in CS 106Z is a random variable with mean of 70. Assume that the exam scores can be real values (i.e. fractional points can be given) but not negative.*

(a) *Give an upper bound for the probability that a student's midterm score will be greater or equal to 80.*

X is a student's midterm score

$$P(X \geq 80) \leq \frac{E[X]}{80} \qquad \text{by Markov's Inequality}$$

$$P(X \geq 80) \leq \frac{7}{8} = .875$$

(b) *Now, say we are given the additional information that the variance of a student's midterm exam score in CS 106Z is 20 (and you can use this information for parts c and d). Give a bound on the probability that a student's midterm score is between 60 and 80, inclusive.*

$$P(60 \leq X \leq 80) = 1 - P(X < 60 \text{ or } X > 80)$$

$$P(|X - \mu| \geq 10) \leq \frac{\sigma^2}{100} \qquad \text{by Chebyshev's Inequality}$$

$$P(60 \leq X \leq 80) \geq 1 - \frac{\sigma^2}{100}$$

$$P(60 \leq X \leq 80) \geq .8$$

(c) *According to Chebyshev's inequality, how many students would have to take the midterm in order to ensure, with at least 90% probability that the class average would be within 5 of 70?*

$$E[\bar{X}] = \mu = 70$$

$$Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{20}{n}$$

$$P(|\bar{X} - E[\bar{X}]| \geq 5) \leq \frac{Var(\bar{X})^2}{25} \qquad \text{by Chebyshev's Inequality}$$

$$P(|\bar{X} - 70| \geq 5) \leq \frac{400}{25n}$$

$$P(|\bar{X} - 70| < 5) > 1 - \frac{400}{25n} \geq .9$$

$$n \geq 160$$

(d) *According to the Central Limit Theorem, how many students would have to take the midterm in order to ensure, with at least 90% probability that the class average would be within 5 of 70?*

$$Z = \frac{(\sum_{i=1}^{n} X_i) - n\mu}{\sigma\sqrt{n}} \text{ as } n \to \infty \qquad \text{by Central Limit Theorem}$$

$$\begin{aligned}
P(-5 \leq \frac{\sum_{i=1}^{n} X_i}{n} - \mu_{\bar{X}} \leq 5) &\geq .9 \\
P(\frac{-5\sqrt{n}}{\sigma} \leq Z \leq \frac{5\sqrt{n}}{\sigma}) &\geq .9 \\
\Phi(\frac{\sqrt{n}}{4}) - \Phi(-\frac{\sqrt{n}}{4}) &\geq .9 \\
\Phi(\frac{\sqrt{n}}{4}) - (1 - \Phi(\frac{\sqrt{n}}{4})) &\geq .9 \\
2\Phi(\frac{\sqrt{n}}{4}) - 1 &\geq .9 \\
n &\geq (4\Phi^{-1}(.95))^2 = 43.29
\end{aligned}$$

4. *Consider a sample of IID exponential random variables $X_1, X_2, ..., X_n$ where each $X_i \sim Exp(\lambda)$*

(a) *Derive the maximum likelihood estimate for the parameter λ in the exponential distribution*

$$\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} f(X_i|\lambda) \\
&= \prod_{i=1}^{n} \lambda e^{-\lambda X_i} \\
LL(\lambda) &= \sum_{i=1}^{n} ln(\lambda e^{-\lambda X_i}) \\
&= \sum_{i=1}^{n} ln(\lambda) - \lambda X_i \\
&= nln(\lambda) - \lambda \sum_{i=1}^{n} X_i \\
\frac{dLL(\lambda)}{d\lambda} &= n\frac{1}{\lambda} - \sum_{i=1}^{n} X_i = 0 \\
\lambda &= \frac{n}{\sum_{i=1}^{n} X_i}
\end{aligned}$$

(b) *Is the estimator you derived in part (a) unbiased?*

No. Unbiased means that if we were to repeat this sampling process many times, the expected value of each estimate would be equal to the true value we are trying to estimate. This can be proved using Jensen's Inequality.

(c) *Is the estimator you derived in part (a) consistent?*

Yes. Consistent means that as the sample size increases, the sampling distribution becomes increasingly concentrated on the true parameter value. This can be proved using the Law of Large Numbers.

5. *Say you have a set of binary input features/variables $X_1, X_2, ..., X_m$ that can be used to make a prediciton about a discrete binary output variable Y (i.e. each of the X_i as well as Y can only take on the values*

0 or 1). In using the input features/variables $X_1, X_2, ..., X_m$ to make a prediction a about Y, recall that the Naïve Bayes classifier makes the simplifying assumption that $P(X_1, X_2, ..., X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$ in order to make it tractable to compute $arg\, max_Y.\ P(X, Y) = arg/, max_Y P(X_1, X_2, ..., X_n|Y)P(Y)$. Say that the first k input variables $X_1, X_2, ..., X_k$ are actually all identical copies of each other, so that when one has the value 0 or 1, they all do. Explain informally, but percisely, why this may be problematic for the model learned by the Naïve Bayes classifier.

This may be problematic because the Naïve Bayes assumption is not true; the inputs are not independent, $P(X_1, X_2, ..., X_n|Y) \neq \prod_{i=1}^{n} P(X_i|Y)$.

6. ***[coding]*** *Implement the Naïve Bayes classifier for the binary input/output data.*

   (a) *After running your algorithms on both the vote and heart data, did you see any difference between the using maximum likelihood estimation versus Laplace estimation in your accuracy results? In general, under what conditions do you think using Laplace estimation would be better? Under what conditions do you think using maximum likelihood estimation would be better?*

   Laplace smoothing generally works better for small datasets, especially when certain cases are missing from the training dataset. Also if Laplace smoothing is used, the program does not need to worry about taking the log of zero. Maximum likelihood estimation is more accurate when the dataset is large and accurately represents the population.