CMPE 343: Introduction to Probability and Statistics for Computer Engineers (Fall 2021)

Homework #1

Due November 29, 2021 by 11:59pm on Moodle

Note: Please type your answers and submit your homework as PDF. To get full points, you need to show your steps clearly. If you use a theorem, rule, definition, derivation, etc. that were not covered in the lectures, you need to cite your resources. If you fail to cite your references, if you plagiarize, if you give your answers to another person, if you copy someone else's answers, your grade will be -100.

1. A function $P : A \subset \Omega \to \mathbb{R}$ is called a **probability law** over the sample space Ω if it satisfies the following three probability axioms.

   - (Nonnegativity) $P(A) \geq 0$, for every **event** A.
   - (Countable additivity) If A and B are two disjoint events, then the probability of their union satisfies

     $$P(A \cup B) = P(A) + P(B).$$

     More generally, for a countable collection of disjoint events $A_1, A_2, ...$ we have

     $$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

   - (Normalization) The probability of the entire sample space is 1, that is, $P(\Omega) = 1$.

   (a) (5 pts) Prove, using only the axioms of probability given, that $P(A) = 1 - P(A^c)$ for any event A and probability law P where A^c denotes the complement of A.

   Solution: We know that by their definition, A and A^c are disjoint and $A \cup A^c = \Omega$
   Therefore, $P(A \cup A^c) = P(A) + P(A^c)$ (Countable Additivity)
   $P(\Omega) = P(A) + P(A^c)$ (From the definition of A and A^c)
   $1 = P(A) + P(A^c)$ (Normalization)
   $P(A) = 1 - P(A^c)$
   Hence, result is proved.

   (b) (5 pts) Let $E_1, E_2, .., E_n$ be disjoint sets such that $\bigcup_{i_1}^{n} E_i = \Omega$ and let P be a probability law over the sample space Ω. Show that, for any event A we have

   $$P(A) = \sum_{i=1}^{n} P(A \cap E_i).$$

Solution:

$$
\begin{aligned}
A &= A \cap \Omega && (1)\\
A &= A \cap \bigcup_{i=1}^{n} E_i && (2)\\
A &= \bigcup_{i=1}^{n} A \cap E_i && (3)\\
P(A) &= P(\bigcup_{i=1}^{n} A \cap E_i) && (4)\\
P(A) &= \sum_{i=1}^{n} P(A \cap E_i) \qquad \text{Countable Additivity} && (5)\\
& && (6)
\end{aligned}
$$

Hence, result is proven.

(c) (5 pts) Prove that for any two events A, B we have

$$P(A \cap B) \geq P(A) + P(B) - 1.$$

Solution: Since any probability must be less than or equal to 1, we can write the inequality below.

$$
\begin{aligned}
P(A \cup B) &= P(A) + P(B) - P(A \cap B) \qquad (AdditiveRule) && (7)\\
1 \geq P(A \cup B) &= P(A) + P(B) - P(A \cap B) && (8)\\
P(A \cap B) &\geq P(A) + P(B) - 1 && (9)
\end{aligned}
$$

Hence, result is proved.

2. (10 pts) Two fair dice are thrown. Let

$$X = \begin{cases} 1, & \text{if the sum of the numbers} \leq 5 \\ 0, & \text{otherwise} \end{cases}$$

$$Y = \begin{cases} 1, & \text{if the product of the numbers is odd} \\ 0, & \text{otherwise} \end{cases}$$

What is $\operatorname{Cov}(X, Y)$? Show your steps clearly.

Solution: We have $6 \cdot 6 = 36$ different cases when we throw a pair of dice. There are only 10 cases whose sum are less than or equal to 5: (1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3),(3,1),(3,2),(4,1)

Realize that only (1,1), (1,3) and (3,1) have odd product when we multiply both numbers. Also, to get an odd number when we multiply two integers, both of them should be odd. Hence, the probability is $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$ Combining these two results, we can write:
P(X=1) = 10/36
P(X=0) = 26/36
P(Y=1) = 1/4
P(Y=0) = 3/4
P(X=1,Y=1) = 1/12

$E[X] = \mu_x = 0 \cdot \frac{26}{36} + 1 \cdot \frac{10}{36} = \frac{5}{18}$
$E[Y] = \mu_y = 0 \cdot \frac{3}{4} + 1 \cdot \frac{1}{4} = \frac{1}{4}$
$E[XY] = \mu_x y = 0 \cdot \frac{11}{12} + 1 \cdot \frac{1}{12} = \frac{1}{12}$

$$Cov(X,Y) = E[XY] - \mu_x \cdot \mu_y$$
$$Cov(X,Y) = \frac{1}{12} - \frac{5}{18} \cdot \frac{1}{4}$$
$$Cov(X,Y) = \frac{6}{72} - \frac{5}{72}$$
$$Cov(X,Y) = \frac{1}{72}$$

3. (10 pts) Derive the mean of Poisson distribution.

Solution: Let X be a Poisson distribution. Then,

$$E(X) = \sum_{x=0}^{\infty} \frac{x \cdot \lambda^x \cdot e^{-\lambda}}{x!} \tag{10}$$
$$E(X) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{(x-1)!} \tag{11}$$

Realize that first term is zero. Hence, we can write as follows.

$$E(X) = e^{-\lambda} \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \tag{12}$$
$$E(X) = e^{-\lambda} \lambda \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \tag{13}$$
$$\tag{14}$$

Now, realize that the summation is is the taylor series of e^{λ} Hence, we can write it as follows.

$$E(X) = e^{-\lambda}\lambda e^{\lambda} \tag{15}$$
$$E(X) = \lambda \tag{16}$$

Hence, result is proved.

4. In this problem, we will explore certain properties of probability distributions and introduce new important concepts.

   (a) (5 pts) Recall Pascal's Identity for combinations: $\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}$
   Use the identity to show the following

   $$(1+x)^N = \textstyle\sum_{m=0}^{N} \binom{N}{m} \cdot x^m$$

   which is called the *binomial theorem*. *Hint*: You can use induction.
   Finally, show that the binomial distribution with parameter p is normalized, that is

   $$\textstyle\sum_{m=0}^{N} \binom{N}{m} \cdot p^m \cdot (1-p)^{(N-m)} = 1$$

   Solution: Let us prove a more generic version of the binomial theorem. Which is

   $$(x+y)^N = \textstyle\sum_{m=0}^{N} \binom{N}{m} \cdot x^m \cdot y^{(N-m)}$$

   Assume N = 1. Then,
   $(x+y)^1 = \sum_{m=0}^{1} \binom{1}{m} x^m \cdot y^{(1-m)}$
   $x + y = x + y$
   Hence, we have a base condition for induction. Now assume that this formula is true for N. We will show that it is true for N+1 as well.

$$(x+y)^{N+1} = (x+y) \cdot \sum_{m=0}^{N} \binom{N}{m} \cdot x^m \cdot y^{(N-m)} \tag{17}$$
$$(x+y)^{N+1} = (x+y) \cdot (\binom{N}{0} y^N + \binom{N}{1} y^{N-1}x^1 + \binom{N}{2} y^{N-2}x^2 + ...) \tag{18}$$
$$(x+y)^{N+1} = \binom{N}{0} y^{N+1} + \binom{N}{0} x^1 y^N + \binom{N}{1} x^1 y^N + \binom{N}{1} x^2 y^{N-1} + \binom{N}{2} x^2 y^{N-1} + \binom{N}{2} x^3 y^{N-2} + .. \tag{19}$$
$$(x+y)^{N+1} = \binom{N+1}{0} y^{N+1} + \binom{N+1}{1} x^1 y^N + \binom{N+1}{2} x^2 y^{N-1} + \binom{N+1}{3} x^3 y^{N-2} + ... \tag{20}$$
$$(x+y)^{N+1} = \sum_{m=0}^{N+1} \binom{N+1}{m} \cdot x^m \cdot y^{(N+1-m)} \tag{21}$$

Hence, we showed that equation holds for N+1 as well. Therefore, this equation is true for all natural numbers N.(It also holds for N = 0, we could have started inductive step from 0 as well) To show the desired result for the first part, just take y = 1.
For the second part of the question, take x = p, y = 1 - p

$$(p+1-p)^N = \sum_{m=0}^{N} \binom{N}{m} \cdot p^m \cdot (1-p)^{(N-m)} \tag{22}$$

$$(1)^N = \sum_{m=0}^{N} \binom{N}{m} \cdot p^m \cdot (1-p)^{(N-m)} \tag{23}$$

$$1 = \sum_{m=0}^{N} \binom{N}{m} \cdot p^m \cdot (1-p)^{(N-m)} \tag{24}$$

Hence, our proof is done.

(b) (5 pts) Suppose you wish to transmit the value of a random variable to a receiver. In Information Theory, the average amount of information you will transmit in the process (in units of "nat") is obtained by taking the expectation of $\ln p(x)$ with respect to the distribution $p(x)$ of your random variable and is given by

$$H(x) = -\int_x p(x) \cdot \ln p(x) \cdot dx$$

This quantity is the *entropy* of your random variable. Calculate and compare the entropies of a uniform random variable $x \sim U(0,1)$ and a Gaussian random variable $z \sim \mathcal{N}(0,1)$. Solution:

$$U(0,1) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

$$H(U(0,1)) = -\int_0^1 1.\ln(1) \;\; dx \tag{25}$$

$$H(U(0,1)) = -\int_0^1 0 \;\; dx \tag{26}$$

$$H(U(0,1)) = 0 \tag{27}$$

$$H(N(0,1)) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\frac{-x^2}{2} \ln(\frac{1}{\sqrt{2\pi}} exp\frac{-x^2}{2}) \;\; dx \tag{28}$$

$$H(N(0,1)) = \frac{-1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} exp\frac{-x^2}{2} (\ln\frac{1}{\sqrt{2\pi}} + \ln{(exp\frac{-x^2}{2})} \;\; dx \tag{29}$$

$$H(N(0,1)) = \frac{-1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \ln\frac{1}{\sqrt{2\pi}} exp\frac{-x^2}{2} \;\; dx \quad - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{-x^2}{2} exp\frac{-x^2}{2} \tag{30}$$

$$H(N(0,1)) = \frac{-1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sqrt{2} \; exp(-y^2) \;\; dy + \frac{1}{2\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 exp\frac{-x^2}{2} \;\; dx \tag{31}$$

$$H(N(0,1)) = -1 \ln\frac{1}{\sqrt{2\pi}} + \frac{1}{2}.\sigma(N(0,1)) \tag{32}$$

$$H(N(0,1)) = \ln\sqrt{2\pi} + \frac{1}{2} \tag{33}$$

(c) In many applications, e.g. in Machine Learning, we wish to approximate some probability distribution using function approximators we have available, for example deep neural networks. This creates the need for a way to measure the similarity or the *distance* between two distributions. One proposed such measure is the *relative entropy* or the Kullback-Leibler divergence. Given two probability distributions p and q the KL-divergence between them is given by

$$KL(p||q) = \int_{-\infty}^{\infty} p(x) \cdot \ln\frac{p(x)}{q(x)} \cdot dx$$

i. (2 pts) Show that the KL-divergence between equal distributions is zero.

Solution: Since we know that p(x) and q(x) are equal, fraction inside logarithm becomes 1. So, we can write the integral as follows.

$$KL(p||q) = \int_{-\infty}^{\infty} p(x). \ln 1 \;\; dx \tag{34}$$

$$KL(p||q) = \int_{-\infty}^{\infty} 0 \;\; dx \tag{35}$$

$$KL(p||q) = 0 \tag{36}$$

ii. (2 pts) Show that the KL-divergence is not symmetric, that is $KL(p||q) \neq KL(q||p)$ in general. You can do this by providing an example.

Solution: Just refer to the result of part iii. It can be seen easily that the function is not symmetric with respect to sigma values, at least due to logarithmic term. Hence we can choose two different normal distribution with different mu and sigma values. Clearly, their KL-divergence will not be symmetric.

iii. (16 pts) Calculate the KL divergence between $p(x) \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $q(x) \sim \mathcal{N}(\mu_2, \sigma_2^2)$ for $\mu_1 = 2, \mu_2 = 1.8, \sigma_1^2 = 1.5, \sigma_2^2 = 0.2$. First, derive a closed form solution depending on $\mu_1, \mu_2, \sigma_1, \sigma_2$. Then, calculate its value. (Only numerical answer without clearly showing your steps will not be graded.)

Remark: We call this measure a *divergence* since a proper *distance* function must be symmetric.

Solution: Just refer to the result of part iii.

$$KL(p||q) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(\frac{-1}{2\sigma_1^2}(s-\mu_1)^2\right) \ln\left(\frac{\frac{1}{\sqrt{2\pi}\sigma_1}\exp\frac{-(x-\mu_1)^2}{2\sigma_1^2}}{\frac{1}{\sqrt{2\pi}\sigma_2}\exp\frac{-(x-\mu_2)^2}{2\sigma_2^2}}\right)\ dx \tag{37}$$

$$KL(p||q) = \frac{1}{\sqrt{2\pi}\sigma_1}\int_{-\infty}^{\infty} \exp\frac{-(x-\mu_1)^2}{2\sigma_1^2}.(ln(\frac{\sigma_2}{\sigma_1}.exp\frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2}))\ dx \tag{38}$$

$$KL(p||q) = \frac{1}{\sqrt{2\pi}\sigma_1}\int_{-\infty}^{\infty} \exp\left(\frac{-1}{2\sigma_1^2}(x-\mu_1)^2\right).(\ln\frac{\sigma_2}{\sigma_1} + \frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2})\ dx \tag{39}$$

$$KL(p||q) = \frac{1}{\sqrt{2\pi}\sigma_1}\int_{-\infty}^{\infty} \exp\left(\frac{-1}{2\sigma_1^2}(x-\mu_1)^2\right)\ln\frac{\sigma_2}{\sigma_1}\ dx- \tag{40}$$

$$\frac{1}{\sqrt{2\pi}\sigma_1}\int_{-\infty}^{\infty} \exp\left(\frac{-1}{2\sigma_1^2}(x-\mu_1)^2\right)\frac{(x-\mu_1)^2}{2\sigma_1^2}\ dx+ \tag{41}$$

$$\frac{1}{\sqrt{2\pi}\sigma_1}\int_{-\infty}^{\infty} \exp\left(\frac{-1}{2\sigma_1^2}(x-\mu_1)^2\right)\frac{(x-\mu_2)^2}{2\sigma_2^2}\ dx \tag{42}$$

$$KL(p||q) = \ln\frac{\sigma 2}{\sigma 1} - \frac{\sigma_1^2}{2\sigma_1^2} + \frac{1}{2\sigma_2^2\sqrt{2\pi}\sigma_1}\int_{-\infty}^{\infty} \exp\left(\frac{-1}{2\sigma_1^2}(x-\mu_1)^2\right)(x-\mu_1+\mu_1-\mu_2)^2\ dx \tag{43}$$

$$KL(p||q) = \ln\frac{\sigma 2}{\sigma 1} - \frac{1}{2} + \frac{1}{2\sigma_2^2\sqrt{2\pi}\sigma_1}\int_{-\infty}^{\infty} (x-\mu_1)^2 \exp\left(\frac{-1}{2\sigma_1^2}(x-\mu_1)^2\right)\ dx+ \tag{44}$$

$$\frac{1}{2\sigma_2^2\sqrt{2\pi}\sigma_1}\int_{-\infty}^{\infty} (\mu_1-\mu_2)^2 \exp\left(\frac{-1}{2\sigma_1^2}(x-\mu_1)^2\right)\ dx+ \tag{45}$$

$$\frac{\mu 1-\mu 2}{\sigma_2^2\sqrt{2\pi}\sigma_1}\int_{-\infty}^{\infty} (x-\mu_1)^2 \exp\left(\frac{-1}{2\sigma_1^2}(x-\mu_1)^2\right)\ dx \tag{46}$$

$$KL(p||q) = \ln\frac{\sigma 2}{\sigma 1} - \frac{1}{2} + \frac{\sigma_1^2}{2\sigma_2^2} + \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2} + 0 \tag{47}$$

$$KL(p||q) = \ln\frac{\sigma 2}{\sigma 1} - \frac{1}{2} + \frac{\sigma_1^2}{2\sigma_2^2} + \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2} \tag{48}$$

If we put the values into the equation, we get 2.34 Hence $KL(p||q) = 2.34$

5. In this problem, we will explore some properties of random variables and in particular that of the Gaussian random variable.

   (a) (7 pts) The convolution of two functions f and g is defined as

$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau$

One can calculate the probability density function of the random variable $Z = X+Y$ using convolution operation with X and Y independent and continuous random variables. In fact,

$f_Z(z) = \int_{-\infty}^{\infty} f_X(\tau)f_Y(z-\tau)d\tau$

Using this fact, find the probability density function of $Z = X + Y$, where X and Y are independent standard Gaussian random variables. Find μ_Z, σ_Z. Which distribution does Z belong to? (Hint: use $\sqrt{\pi} = \int_{-\infty}^{\infty} e^{-x^2} \mathrm{d}x$)

Solution:

$$f_z(z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-z)^2}{2}} \; dx \tag{49}$$

$$f_z(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{(-x^2+zx-\frac{z^2}{2})} \; dx \tag{50}$$

$$f_z(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{(-(x-\frac{z}{2})^2-\frac{z^2}{4})} \; dx \tag{51}$$

$$\tag{52}$$

Let $u = x - \frac{z}{2}$ and $du = dx$

$$f_z(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{(-u^2-\frac{z^2}{4})} \; du \tag{53}$$

$$f_z(z) = \frac{e^{\frac{-z^2}{4}}}{2\pi} \int_{-\infty}^{\infty} e^{(-u^2)} \; du \tag{54}$$

$$f_z(z) = \frac{e^{\frac{-z^2}{4}}}{2\pi} \cdot \sqrt{\pi} \tag{55}$$

$$f_z(z) = \frac{e^{\frac{-z^2}{4}}}{2\sqrt{\pi}} \tag{56}$$

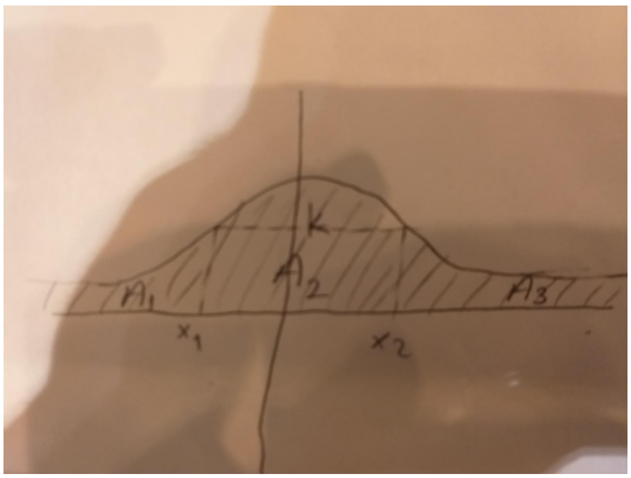
(b) (5 pts) Let X be a standard normal Gaussian random variable and Y be a discrete random variable taking values $\{-1, 1\}$ with equal probabilities. Is the random variable $Z = XY$ independent of Y? Give a formal argument(proof or counter example) justifying your answer.

Warning: My solution for this question is a bit long.
Solution: For Z and Y to be dependent, knowing the value of Y or Z should give us information about the other one. Depending on how Y is defined, Z and Y might either dependent or independent. Hence, we cannot reach to a certain conclusion without further information about Y. I will show that by giving 2 different definitions of Y and reaching dependency in one of them and independency in the other one:

Case1: Y gets its value by tossing a fair coin. Head corresponds to 1, tail corresponds to -1
Either Y = 1 or -1, Z = XY equals to Z = X, since Y doesn't depend on X, we cannot derive any information for the value of Z. Hence, they must be independent.

Case2: Y gets its value from X. Consider the function Y below:

$$Y = \begin{cases} 1, & \text{X(x)} > \text{k} \\ -1, & \text{X(x)} \leq \text{k} \end{cases}$$



There must be a value k which divides the function into three are as shown such that $A_1 = A_3$ and $A_1 + A_3 = A_2$ (I didn't calculate the value of k but there must be such) Thus, Y = 1 and Y = -1 are equally likely. Now, consider that we know the value of Y. Then, we can immediately say that Z is either less than or greater than k. Hence, they are partially dependent.

By these two examples, we can conclude that without having the further information about Y, we cannot conclude whether Y and Z are dependent or independent.

(c) (8 pts) Let X be a non-negative random variable. Let k be a positive real number. Define the binary random variable $Y = 0$ for $X < k$ and $Y = k$ for $X \geq k$. Using the relation between X and Y, prove that $P(X \geq k) \leq \frac{E[X]}{k}$. (Hint: start with finding $E[Y]$).

Solution:
From the definition of X and Y, we can see that for all values of x, value of X is greater than Y. Using this observation, we can write:

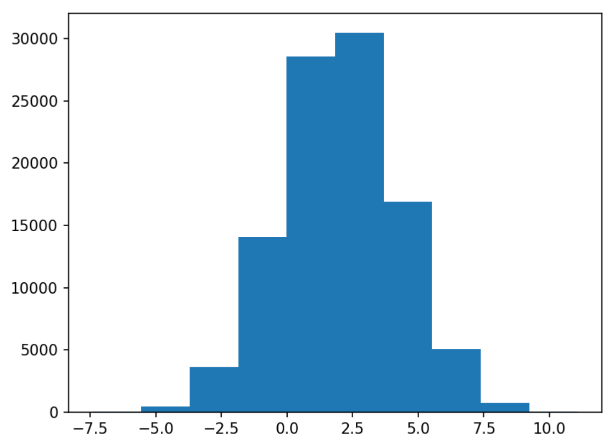
$$E[Y] = P(X \geq k) \cdot k \leq E[X] \tag{57}$$

$$\frac{E[Y]}{k} = P(X \geq k) \leq \frac{E[X]}{k} \tag{58}$$

Hence, our proof is done.

6. In this problem, we will empirically observe some of the results we obtained above and also the convergence properties of certain distributions. You may use the python libraries Numpy and Matplotlib.

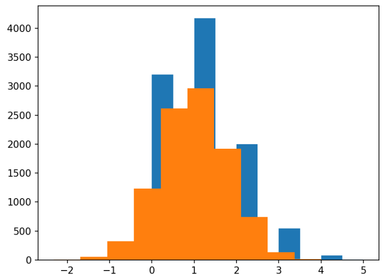
   (a) (5 pts) In *3.a* you have found the distribution of Z = X+ Y. Let X and Y be Gaussian random variables with $\mu_X = -1$, $\mu_Y = 3$, and $\sigma_X^2 = 1$ and $\sigma_Y^2 = 4$. Sample 100000 pairs of X and Y and plot their sum Z = X+Y as a histogram. Is the shape of Z and its apparent mean consistent with *what you have learned in the lectures*?

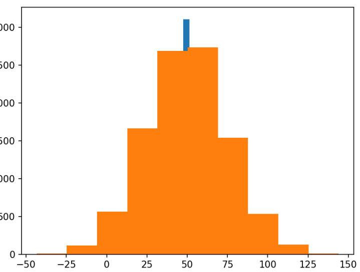


Solution: Yes, they are consistent. The distribution is equal. The mean of Z should be -1 + 3 = 2 and the distribution should look like a normal distribution as expected and histogram confirms this.

   (b) (5 pts) Let $X\ B(n, p)$ be a binomially distributed random variable. One can use the normal distribution as an approximation to the binomial distribution when n is large and/or p is close to 0.5. In this case, $X \approx N(np, np(1-p))$. Show how such approximation behaves by drawing 10000 samples from binomial distribution with $n = 5, 10, 20, 30, 40, 100$ and $p = 0.2, p = 0.33, 0.50$ and plotting the distributions of samples for each case as a histogram. Report for which values of n and p the distribution resembles that of a Gaussian?

Solution: Histograms are similar to as expected. As n increases and p becomes closer to 0.5, approximation becomes better. I added two edge case screenshots where in the first one n= 5 and p = 0.2, and in the second one n = 100 and p = 0.5.





(c) (5 pts) You were introduced to the concept of KL-divergence analytically. Now, you will estimate this divergence $KL(p||q)$. Where $p(x) = \mathcal{N}(0, 1)$ and $q(x) = \mathcal{N}(0, 4)$. Sample 1000 samples from a Gaussian with mean 0 and variance 1. Call them $x_1, x_2, ..., x_{1000}$. Estimate the KL divergence as

$$\frac{1}{1000}\sum_{i=1}^{1000} \ln \frac{p(x_i)}{q(x_i)}$$

where $p(x) = \mathcal{N}(0, 1)$ and $q(x) = \mathcal{N}(0, 4)$. Calculate the divergence analytically for $KL(p||q)$. Is the result consistent with your estimate?

Solution: According to the formula we have derived from the question 4.iii, just subtitute $\sigma_1 = 1$ $\sigma_2 = 2$ Then, theoric analysis gives us approximately 0.318 My code also produces similar values as to theoric result.