

Sundry

Before you start your homework, write down your team. Who else did you work with on this homework? List names and email addresses. (In case of homework party, you can also just describe the group.) How did you work on this homework? Working in groups of 3-5 will earn credit for your "Sundry" grade.

Please copy the following statement and sign next to it:

I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.

1 Whitening

Let X and Y be two random variables, with $\text{var}(X) > 0, \text{var}(Y) > 0$. Show that it is possible to construct $\tilde{X} = aX + bY$ and $\tilde{Y} = cX + dY$, where a, b, c, d are scalars to be chosen subject to the constraint $ad - bc \neq 0$, such that $\text{cov}(\tilde{X}, \tilde{Y}) = 0$.

2 Will I Get My Package?

Sneaky delivery guy of some company is out delivering n packages to n customers. Not only does he hand a random package to each customer, he opens the package before delivering it with probability $1/2$. Let X be the number of customers who receive their own packages unopened.

- (a) Compute the expectation $\mathbb{E}(X)$.
- (b) Compute the variance $\text{var}(X)$.

3 Student Request Collector

After a long night of debugging, Alvin has just perfected the new homework party/office hour queue system. CS 70 students sign themselves up for the queue, and TAs go through the queue,

resolving requests one by one. Unfortunately, our newest TA (let's call him TA Bob) does not understand how to use the new queue: instead of resolving the requests in order, he always uses the Random Student button, which (as the name suggests) chooses a random student in the queue for him. To make matters worse, after helping the student, Bob forgets to click the Resolve button, so the student still remains in the queue! For this problem, assume that there are n total students in the queue.

- (a) Suppose that Bob has already helped k students. What is the probability that the Random Student button will take him to a student who has not already been helped?
- (b) Let X_i^r be the event that TA Bob has not helped student i after pressing the Random Student button a total of r times. What is $\mathbb{P}[X_i^r]$? Assume that the results of the Random Student button are independent of each other. Now approximate the answer using the inequality $1 - x \leq e^{-x}$.
- (c) Let T_r represent the event that TA Bob presses the Random Student button r times, but still has not been able to help all n students. (In other words, it takes TA Bob longer than r Random Student button presses before he manages to help every student). What is T_r in terms of the events X_i^r ? [*Hint*: Events are subsets of the probability space Ω , so you should be thinking of set operations. . .]
- (d) Using your answer for the previous part, what is an upper bound for $\mathbb{P}[T_r]$? (You may leave your answer in terms of your approximation in part (b).)
- (e) Now let $r = \alpha n \ln n$. What is an upper bound for $\mathbb{P}[X_i^r]$?
- (f) Calculate an upper bound for $\mathbb{P}[T_r]$ using the same value of r as before. (This is more formally known as a bound on the tail probability of the distribution of button presses required to help every student. This distribution will be explored in more detail later, in the context of random variables.)
- (g) What value of r do you need to bound the tail probability by $1/n^2$? In other words, how many button presses are needed so that the probability that TA Bob has not helped every student is at most $1/n^2$?

4 Student Life

In an attempt to avoid having to do laundry often, Marcus comes up with a system. He owns 10 shirts that are stored in his dresser. Every night, he designates his dirtiest shirt and throws it on his chair. In the morning, he randomly picks one of his shirts to wear (the ones in the dresser and the one on the chair). If he picked up the dirtiest one, he puts it in the dirty pile at the end of the day, and designates one of his remaining shirts as the current dirtiest one. He does his laundry once he's out of shirts.

- (a) What is the expected number of days it takes for him to do his laundry?

- (b) Say he gets even lazier, and instead of organizing his shirts as he had planned, on each day he throws his shirts randomly into one of 10 different locations in his room (one shirt per location). Each day, he designates one of his shirts as his dirtiest shirt and one location as the dirtiest location. In the morning, if he happens to pick the dirtiest shirt, *and* the dirtiest shirt was in the dirtiest location, then he puts the shirt into the dirty pile and does not use that location anymore (it is too dirty now). What is the expected number of days for him to do his laundry now?

5 Darts with Friends

Sayan and Sinho are playing darts. Being the better player, Sinho's aim follows a uniform distribution over a circle of radius r around the center. Sayan's aim follows a uniform distribution over a circle of radius $2r$ around the center.

- (a) Let the distance of Sinho's throw be denoted by the random variable X and let the distance of Sayan's throw be denoted by the random variable Y .
- What's the cumulative distribution function of X ?
 - What's the cumulative distribution function of Y ?
 - What's the probability density function of X ?
 - What's the probability density function of Y ?
- (b) What's the probability that Sinho's throw is closer to the center than Sayan's throw? What's the probability that Sayan's throw is closer to the center?
- (c) What's the cumulative distribution function of $U = \min\{X, Y\}$?
- (d) What's the cumulative distribution function of $V = \max\{X, Y\}$?
- (e) What is the expectation of the absolute difference between Sinho's and Sayan's distances from the center, that is, what is $\mathbb{E}[|X - Y|]$? [*Hint*: There are two ways of solving this part.]

6 Exponential Practice

Let $X \sim \text{Exponential}(\lambda_X)$ and $Y \sim \text{Exponential}(\lambda_Y)$ be independent, where $\lambda_X, \lambda_Y > 0$. Let $U = \min\{X, Y\}$, $V = \max\{X, Y\}$, and $W = V - U$.

- (a) Compute $\mathbb{P}(U > t, X \leq Y)$, for $t \geq 0$.
- (b) Use the previous part to compute $\mathbb{P}(X \leq Y)$. Conclude that the events $\{U > t\}$ and $\{X \leq Y\}$ are independent.
- (c) Compute $\mathbb{P}(W > t \mid X \leq Y)$.
- (d) Use the previous part to compute $\mathbb{P}(W > t)$.
- (e) Calculate $\mathbb{P}(U > u, W > w)$. Conclude that U and W are independent. [*Hint*: Think about the approach you used for the previous parts.]

7 Noisy Love

Due to the Central Limit Theorem, the Gaussian distribution is often used as a model for noise. In this problem, we will see how to perform calculations with Gaussian noise models.

Suppose you have confessed to your love interest on Valentine's Day and you are waiting to hear back. Your love interest is trying to send you a binary message: "0" means that your love interest is not interested in you, while "1" means that your love interest reciprocates your feelings. Let X be your love interest's message for you. Your current best guess of X has $\mathbb{P}(X = 0) = 0.7$ and $\mathbb{P}(X = 1) = 0.3$. Unfortunately, your love interest sends you the message through a noisy channel, and instead of receiving the message X , you receive the message $Y = X + \varepsilon$, where ε is independent Gaussian noise with mean 0 and variance 0.49.

- First, you decide upon the following rule: if you observe $Y > 0.5$, then you will assume that your love interest loves you back, whereas if you observe $Y \leq 0.5$, then you will assume that your love interest is not interested in you. What is the probability that you are correct using this rule? (Express your answer in terms of the CDF of the standard Gaussian distribution $\Phi(z) = \mathbb{P}(\mathcal{N}(0, 1) \leq z)$, and then evaluate your answer numerically.)
- Suppose you observe $Y = 0.6$. What is the probability that your love interest loves you back? [*Hint*: This problem requires conditioning on an event of probability 0, namely, the event $\{Y = 0.6\}$. To tackle this problem, think about conditioning on the event $\{Y \in [0.6, 0.6 + \delta]\}$, where $\delta > 0$ is small, so that $f_Y(0.6) \cdot \delta \approx \mathbb{P}(Y \in [0.6, 0.6 + \delta])$, and then apply Bayes Rule.]
- Suppose you observe $Y = y$. For what values is it more likely than not that your love interest loves you back? [*Hint*: As before, instead of considering $\{Y = y\}$, you can consider the event $\{Y \in [y, y + \delta]\}$ for small $\delta > 0$. So, when is $\mathbb{P}(X = 1 \mid Y \in [y, y + \delta]) \geq \mathbb{P}(X = 0 \mid Y \in [y, y + \delta])$?
- Your new rule is to assume that your love interest loves you back if (based on the value of Y that you observe) it is more likely than not that your love interest loves you back. Under this new rule, what is the probability that you are correct?

8 Binomial CLT

In this question we will explicitly see why the central limit theorem holds for the binomial distribution as the number of coin tosses grows.

Let X be the random variable showing the total number of heads in n independent coin tosses.

- Compute the mean and variance of X . Show that $\mu = \mathbb{E}[X] = n/2$ and $\sigma^2 = \text{var} X = n/4$.
- Prove that $\mathbb{P}[X = k] = \binom{n}{k}/2^n$.
- Show by using Stirling's formula that

$$\mathbb{P}[X = k] \simeq \frac{1}{\sqrt{2\pi}} \left(\frac{n}{2k}\right)^k \left(\frac{n}{2(n-k)}\right)^{n-k} \sqrt{\frac{n}{k(n-k)}}.$$

In general we expect $2k$ and $2(n-k)$ to be close to n for the probability to be non-negligible. When this happens we expect $\sqrt{\frac{n}{k(n-k)}}$ to be close to $\sqrt{\frac{n}{(n/2) \times (n/2)}} = 2/\sqrt{n}$. So replace that part of the formula by $2/\sqrt{n}$.

- (d) In order to normalize X , we need to subtract the mean, and divide by the standard deviation. Let $Y = (X - \mu)/\sigma$ be the normalized version of X . Note that Y is a discrete random variable. Determine the set of values that Y can take. What is the distance d between two consecutive values?
- (e) Let $X = k$ correspond to the event $Y = t$. Then $X \in [k - 0.5, k + 0.5]$ corresponds to $Y \in [t - d/2, t + d/2]$. For conceptual simplicity, it is reasonable to assume that the mass at point t is distributed uniformly on the interval $[t - d/2, t + d/2]$. We can capture this with the idea of a “probability density” and say that the probability density on this interval is just $\mathbb{P}[Y = t]/d = \mathbb{P}[X = k]/d$.

Compute k as a function of t . Then substitute that for k in the approximation you have from part c to find an approximation for $\mathbb{P}[Y = t]/d$. Show that the end result is equivalent to:

$$\frac{1}{\sqrt{2\pi}} \left(\left(1 + \frac{t}{\sqrt{n}}\right)^{1+t/\sqrt{n}} \left(1 - \frac{t}{\sqrt{n}}\right)^{1-t/\sqrt{n}} \right)^{-n/2}$$

- (f) As you can see, we have expressions of the form $(1+x)^{1+x}$ in our approximation. To simplify them, write $(1+x)^{1+x}$ as $\exp((1+x)\ln(1+x))$ and then replace $(1+x)\ln(1+x)$ by its Taylor series.

The Taylor series up to the x^2 term is $(1+x)\ln(1+x) \simeq x + x^2/2 + \dots$ (feel free to verify this by hand). Use this to simplify the approximation from the last part. In the end you should get the familiar formula that appears inside the CLT:

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right).$$

(The CLT is essentially taking a sum with lots of tiny slices and approximating it by an integral of this function. Because the slices are tiny, dropping all the higher-order terms in the Taylor expansion is justified.)