Brain Teasers

Problem. (Difficulty: Easy) There are N lions and 1 sheep in a field. Every lion wants primarily to not get eaten, and secondly to eat a sheep. The catch is that if a lion eats a sheep, he becomes a sheep himself. What will happen in the field if every lion is a perfect logician? E.g., if $N = 1$, the lion will eat the sheep. But if $N = 2$, neither lion would eat the sheep because then he would be eaten by the other lion.

Solution. Use induction to find that if N is even no lion would eat the sheep, but if N is odd each lion woul try to immediately eat the sheep.

Problem. (Difficulty: Easy) A horizontal stick is one meter long. Fifty ants are placed in random positions on the stick, pointing in random directions. The ants crawl head first along the stick, moving at one meter per minute. If an ant reaches the end of the stick, it falls off. If two ants meet, they both change direction. How long do you have to wait to be sure that all the ants have fallen off the stick?

Solution. Instead of bouncing, pretend ants that meet move through each other. Then it is clear that the answer is $\boxed{\text{1 minute}}$.

Problem. (Difficulty: Hard) You have 12 balls, 11 of which are identical. The 12th ball has a weight distinct from the other 11. Given a balance scale, what is the least number of uses of the scale you need to determine the unique ball?

Solution. In general, you can distinguish the unique ball from a set of $\frac{3^n - 3}{2}$ balls with n uses of the scale. The idea is based on the easier problem of finding the unique ball from a group of 3^n balls if you know whether the unique ball is heavier or lighter (the algorithm for this is easy - keep splitting evenly into groups of three and weighing two of the groups against one another).

The algorithm then goes as follows: split the $\frac{3^n - 3}{2}$ balls into three groups of size $\frac{3^{n-1} - 1}{2}$ balls, and then split each of those groups into subgroups of size $3^{n-2}, 3^{n-1}, \ldots, 1$. For the first step, weigh two of the main groups against one another, and then rotate the subgroups of size 3^{n-2} among the three groups. If the result of the weighing changed, then the unique ball is in an identifiable subgroup of size 3^{n-2} and you know whether it is heavier or lighter based on the two weighings you've already done. Then you can finish by finding the unique ball in this set of 3^{n-2} balls with $n - 2$ weighings as described in the first paragraph. Otherwise, rotate the subgroups of 3^{n-3} among the three groups and use the same logic, and so on. This finishes the argument.

Statistics

Problem. (Difficulty: Easy) Give an example of two variables that are dependent but uncorrelated.

Solution. Take $X \sim \text{Unif}(-1, 1)$ and $Y = X^2$. Clearly they are dependent, but it is easy to see that they have correlation 0.

Problem. (Difficulty: Easy) Given a variable X, construct a variable Y that has a given correlation r to X.

Solution. Assume WLOG that $Var(X) = 1$. Then let $\boxed{Y = rX + \sqrt{1 - r^2} \cdot N(0, 1)}$. It is easy to verify that $Var(Y) = 1$ and $Cov(X, Y) = rVar(X) = r$ as desired.

Problem. (Difficulty: Easy) Your friend Emily claims she can tell differently colored skittles apart by taste. There are five colors of skittles (each equally likely to appear). You give her three skittles and she guessing the color correctly twice. Do you believe her claim? What if you gave her 100 skittles and she guessed the color correctly 40 times?

Solution. The null hypothesis here is that she can't tell them apart, so she has a 0.2 chance of guessing correctly for each skittle. The probability she gets two out of three right is then equal to $3(0.2)^2(0.8) = 0.096$, which is low but not convincing.

To calculate the probability she gets 40 out of 100 right, we use the fact that the results of this set of 100 Bernoulli trials can be modeled by a binomial distribution, which is well-approximated by a normal distribution. In this case the mean of the normal distribution is 20 and the standard deviation is $\sqrt{100 \cdot 0.2 \cdot (1-0.2)} = 4$. Thus the z-score for Emily getting 40 skittles right is $\frac{40-20}{4} = 5$, which corresponds to a miniscule probability that the null hypothesis is true. Thus in this case, her performance is convincing.

Problem. (Difficulty: Easy) Suppose you have three variables X, Y, Z and you know that $corr(X,Y) = 0.6$ and $corr(Y,Z) = 0.8$. What are the possible values for $corr(X,Z)$?

Solution. One solution is to interpret correlations as cosines. Suppose the "angle" between X and Y is α and the "angle" between Y and Z is β. Then the angle between X and Z must be between $\alpha - \beta$ and $\alpha + \beta$. Thus $corr(X,Z)$ must be between

$$\cos(\alpha-\beta) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta) = 0.6 \cdot 0.8 + \sqrt{1-0.6^2} \cdot \sqrt{1-0.8^2} = 0.96$$

and

$$\cos(\alpha-\beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta) = 0.6 \cdot 0.8 - \sqrt{1-0.6^2} \cdot \sqrt{1-0.8^2} = 0$$

More generally, being *positive semidefinite* (having all eigenvalues non-negative) is a necessary and sufficient condition for being a valid correlation matrix. Thus we want the values of r for which the matrix

$$\begin{pmatrix} 1 & 0.6 & r \\ 0.6 & 1 & 0.8 \\ r & 0.8 & 1 \end{pmatrix}$$

is positive semidefinite.

By Sylvester's criterion, a matrix is positive semidefinite if and only if the upper right hand $k \times k$ sub-matrix of the matrix has positive determinant for all k. Thus in this case we only need to check the determinant of the full matrix, which is equal to $0.96r - r^2$ from which we immediately get that $r \in [0, 0.96]$ as before.

Problem. (Difficulty: Medium) Suppose you have two series of data X and Y and when you run two simple linear regressions between X and Y you obtain the relationships $Y = aX + b$ and $X = cY + d$. Find the range of possible values for the product ac.

Solution. It is well-known that $a = \frac{Cov(X,Y)}{Var(X)}$ and $c = \frac{Cov(X,Y)}{Var(Y)}$. Thus $ac = r^2$ where r is the correlation between X and Y, which can clearly take on any value in $[0,1]$.

Problem. (Difficulty: Medium) Given i.i.d. random variables $X, Y \sim N(0,1)$, find $P(X = x | X+Y > 0)$ and prove that your result corresponds to a valid probability distribution. Express your answer in terms of *pdf* and *cdf*, the probability and cumulative density functions for $N(0,1)$.

Solution. By Bayes' rule we have that

$$P(X = x | X+Y > 0) = \frac{P(X+Y>0|X=x)P(X=x)}{P(X+Y>0)} = \boxed{2cdf(x)pdf(x)}$$

To prove that this is a valid probability distribution we need to show that

$$2\int_{-\infty}^{\infty} cdf(x)pdf(x)\ dx = 1$$

The key idea here is that $pdf(x) = pdf(-x)$ and that $cdf(-x) + cdf(x) = 1$ for all x, both of which follow from the fact that $N(0,1)$ is symmetric around 0. Thus we can write

$$2\int_{-\infty}^{\infty} cdf(x)pdf(x)\ dx = 2\int_{0}^{\infty} cdf(-x)pdf(-x)\ dx + 2\int_{0}^{\infty} cdf(x)pdf(x)\ dx = 2\int_{0}^{\infty} pdf(x)\ dx = 1$$

as desired.

Problem. (Difficulty: Medium) Suppose you draw n samples $x_1, x_2, \ldots, x_n$ from a distribution. Using these samples, compute un-biased estimators for the mean μ and variance σ^2 of the population.

Solution. Let $m = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $s^2 = \frac{1}{n}\sum_{i=1}^{n}(m - x_i)^2$ be the sample mean and variance. Then we have that

$$\mathbb{E}[m] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

so $\boxed{m}$ is an unbiased estimator for μ.

Now note that since m is an average of n samples,

$$\mathbb{E}\left[(\mu - m)^2\right] = \frac{1}{n}\cdot\mathbb{E}\left[(\mu - x_i)^2\right] = \frac{\sigma^2}{n}$$

But also note that

$$\mathbb{E}\left[(\mu - x_i)^2\right] = \mathbb{E}\left[(m - x_i)^2\right] + \mathbb{E}\left[(\mu - m)^2\right]$$

so

$$\sigma^2 = \mathbb{E}[s^2] + \frac{\sigma^2}{n}$$

Thus $\boxed{\frac{n}{n-1}\cdot s^2}$ is an unbiased estimator for σ^2. The intuition for why we need to increase s^2 to get an unbiased estimator for σ^2 is that s^2 comes from the sum of squares of the distance between m and all of the samples. But m is the value that minimizes the sum of squares - in actuality, the sum of squares using the actual mean of the population should be higher. The $\frac{n}{n-1}$ coefficient is known as the *Bessel correction.*

Problem. (Difficulty: Easy, Easy, Easy, Hard) Suppose X, Y are i.i.d random variables drawn from $N(0,1)$.

(a) Compute $P(X > Y)$

(b) Compute $P(X > 2Y)$

(c) Compute $P(X > |Y|)$

(d) Compute $P(X > 2|Y|)$

Solution. For part (a), X and Y are symmetric so the probability is $\boxed{0.5}$. For part (b), $X - 2Y$ is a normal distribution with mean 0 and standard deviation $\sqrt{5}$. Since this is symmetric around 0, the probability is $\boxed{0.5}$. For part (c), split into two cases:

$$P(X > |Y|) = P(X \geq 0)P(X > |Y||X \geq 0) + P(X < 0)P(X > |Y||X < 0) = 0.5 \cdot 0.5 + 0.5 \cdot 0 = \boxed{0.25}$$

Part (d) is a little trickier. The idea is to consider the joint distribution (X, Y) on a plane and note that the probability distribution has circular symmetry around the origin. The part of the plane for which $X > 2|Y|$ is the sector expanding to the right between lines $x = 2y$ and $x = -2y$. The angle of

the sector is $2\tan^{-1}(0.5)$ and since we have circular symmetry the probability that a randomly chosen point from the joint distribution (X, Y) lies in this sector is given by $\frac{2\tan^{-1}(0.5)}{2\pi} = \boxed{\frac{\tan^{-1}(0.5)}{\pi}}$.

Problem. (Difficulty: Hard) Suppose X is log-normal, so that $\log X \sim N(\mu, \sigma)$. What is $\mathbb{E}[X]$?

Solution. Using a change of variables, we can see that the density function for X is given by

$$P(X = x) = \frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

Then

$$\mathbb{E}[X] = \frac{1}{\sqrt{2\pi}\sigma} \int_0^\infty e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}\, dx$$

Let $z = \frac{\ln x - \mu}{\sigma}$ so that $dx = \sigma x dz = \sigma e^{\mu + \sigma z} dz$. Then we have

$$\frac{1}{\sqrt{2\pi}\sigma} \int_0^\infty e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{z^2}{2} + \sigma z + \mu}\, dz = \frac{e^{\mu + \frac{\sigma^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{(z-\sigma)^2}{2}}\, dz = \boxed{e^{\mu + \frac{\sigma^2}{2}}}$$

Where for the last step we used the fact that the integral over the pdf of the normal distribution $N(\sigma, 1)$ is 1.

Probability/Expected Value

Problem. (Difficulty: Easy) What is the expected number of draws from a standard deck of cards until you first see an ace?

Solution. Note that the four aces split the deck into five sections, each with an average of $\frac{52-4}{5} = 9.6$ cards. Thus the expected number of draws till an ace is seen is $9.6 + 1 = \boxed{10.6}$.

Problem. (Difficulty: Easy) Every second, a single-celled organism either splits into two cells (probability 0.3), does nothing (probability 0.5), or dies (probability 0.2). What is the probability that the organism goes extinct?

Solution. Let p denote the probability of the organism going extinct. Then we can write the recurrence $p = 0.3p^2 + 0.5p + 0.2$ which has solutions $p = 1$ and $\boxed{p = 2/3}$.

Problem. (Difficulty: Easy) Find the eigenvalues of an $n \times n$ matrix M with 1s on the main diagonal and r everywhere else.

Solution. There are multiple ways to do this problem. The simplest is to write the matrix as $M = rE_n + (1-r)I_n$ where E_n is the $n \times n$ matrix of all 1s and I_n is the $n \times n$ identity matrix. The eigenvalues of M are $1 - r$ added to each of the the eigenvalues of rE_n. Note that rE_n has rank 1 and so $n - 1$ of its eigenvalues are 0. Its trace is nr, so the remaining eigenvalue is nr. Thus the eigenvalues of M are $n - 1$ eigenvalues of $\boxed{1 - r}$, and one eigenvalue of $\boxed{nr + 1 - r}$.

Problem. (Difficulty: Easy) Suppose I flip a fair coin until I see the sequence HHT. What is the expected number of flips I will make?

Solution. Let X denote this expected number. Let X_H denote the expected number of flips remaining if we just saw a tails. And let X_{HH} denote the expected number of flips yet to make if we just saw a

two heads in a row. Then, conditioning on the result of the "next flip" we can write the equations

$$\begin{aligned} X &= 0.5(X_H + 1) + 0.5(X + 1) \\ X_H &= 0.5(X_{HH} + 1) + 0.5(X + 1) \\ X_{HH} &= 0.5(X_{HH} + 1) + 0.5(3) \end{aligned}$$

which is a system of three-variables we can easily solve to obtain $\boxed{X = 8}$.

Problem. (Difficulty: Easy) Suppose I flip a fair coin until I see n heads in a row. What is the expected number of flips I will make?

Solution. Let X denote this expected number. Then if we flip $k < n$ heads in a row and then a tails, our expected number of flips will be $k + 1 + X$ and the probability that this occurs is 2^{-k-1}. There is also a 2^{-n} probability that we immediately flip n heads in a row. Thus we can write the equation

$$X = 2^{-n} \cdot n + \sum_{k=0}^{n-1} 2^{-k-1} \cdot (X + k + 1)$$

This is merely a linear equation in X which we can solve to get $\boxed{X = 2^{n+1} - 2}$.

Problem. (Difficulty: Easy) Isabella plays a game where she rolls a dice up to three times, but for any roll can decide to keep whatever number it landed on (untl the third, where she must keep it). She wants to keep as high a number as possible. With optimal play, what is the expected value of the number she keeps.

Solution. If she were playing with one roll instead of three, the expected value would be $7/2$. If she were playing with two rolls, based on the expected value of the one roll game she would only choose to re-roll her first roll if it was under $7/2$. Thus there is a $1/2$ chance her first roll is a $4, 5$, or 6 in which case the expected value is 5, and a $1/2$ chance she re-rollsand move to the one roll game, which has an expected value of $7/2$. Therefore the expected value of the two roll game is $1/2 * 7/2 + 1/2 * 5 = 17/4$. Now in the actual three roll game, based on the expected value of the two roll game, she will only keep her first roll if it is a 5 or a 6, and otherwise she will re-roll and move to the two-roll game. Thus the expected value of the game is $1/3 * 11/2 + 2/3 * 17/4 = \boxed{14/3}$.

Problem. (Difficulty: Easy) Emily and Isabella are playing a game where Emily picks a random real number from the interval $[0, 1]$, and then Isabella picks random real numbers from the interval $[0, 1]$ until her number is smaller than Emily's. What is the expected number of times Isabella will pick a number until the game finishes?

Solution. If Emily picked $x \in [0, 1]$ then the probability Isabella picks a number smaller than x is x. Thus the expected number of times she has to pick is $\dfrac{1}{x}$. Thus the answer is given by

$$\int_0^1 \frac{1}{x}\, dx = \boxed{\infty}$$

Problem. (Difficulty: Easy) What is the expected number of die rolls needed before seeing every possible number appear?

Solution. Suppose we have already seen k distinct numbers. Then the probability we see a new number on our next roll is given by $\dfrac{6-k}{6}$, so by geometric probability the expected number of rolls until we see a new number is $\dfrac{6}{6-k}$. Thus the answer is

$$\frac{6}{6} + \frac{6}{5} + \cdots + \frac{6}{1} = \boxed{6H_6}$$

where H_n denotes the nth harmonic number. This problem is known as the *Coupon Collector's Problem*.

Problem. (Difficulty: Easy). Suppose n cars are on a one-lane highway in some order, and that each car moves at some randomly speed chosen uniformly from $[0, 1]$. If a car is faster than a car directly in front of it, it will form a "clump" with the car in front of it and begin to move at the same speed. What is the expected number of clumps?

Solution. Let $f(n)$ denote the expected number of clumps. Consider the position of the fastest car. If it is the front-most car then it will not form a clump with any other car, and so the expected number of clumps will be $f(n-1)+1$. Otherwise, it will form a clump with the car in front of it which we can treat as one car. In this case, the expected number of clumps is just $f(n-1)$. Therefore we can write the recurrence

$$f(n) = \frac{1}{n} \cdot (f(n-1)+1) + \frac{n-1}{n} \cdot f(n-1) = f(n-1) + \frac{1}{n}$$

This immediately leads to the answer $f(n) = \boxed{H_n}$.

Problem. (Difficulty: Easy) You are given n unit vectors in n-dimensional space. Find a unit vector that forms the same angle with each of these n vectors.

Solution. Label these N vectors $v_1, v_2, \ldots v_n$ and suppose for all i that $v_i = (a_{i,1}, a_{i,2}, \ldots, a_{i,n})$. Suppose the vector we want is given by $x = (x_1, x_2, \ldots x_n)$. It is well-known that the dot product of two unit vectors is equal to the cosine between them, so we want to find the x so that $v_1 \cdot x = v_2 \cdot x = \cdots = v_n \cdot x = 1$. But if let $A = [a_{ij}]$ denote the matrix of as, this is equivalent to

$$Ax = 1$$

where 1 denotes the $n \times 1$ vector of 1s. Then we immediately have $\boxed{x = A^{-1}1}$

A more efficient solution might be to notice that $x \cdot (v_i - v_1) = 0$ for all $i \in \{2, 3, \ldots, n\}$ and to use Gram-Schmidt.

Problem. (Difficulty: Medium) Isabella is playing a game in which she throws darts at a dartboard until her most recent dart is further away from the bullseye than any dart she threw previously. What is the expected number of darts Isabella throws before the game ends?

Solution. Suppose the game ends with Isabella having thrown $k \geq 2$ darts. Label the k darts with the integers $1, 2, \ldots, k$ where the dot with label i is the ith closest from the bullseye. Consider all the possible orders in which Isabella threw the darts. The $k-1$th dart must have label 1, and the kth dart can have any label. Once the kth dart's label is fixed, the remaining $k-2$ darts must have been thrown in decreasing order of label. Thus there are $k-1$ ways she could have ordered her throws, among $k!$ possible orderings in total. Thus the probability the game ends after k throws is $\dfrac{k-1}{k!}$, and so the expected number of darts thrown is

$$\sum_{k=2}^{\infty} \frac{k-1}{k!} \cdot k = \sum_{k=2}^{\infty} \frac{1}{(k-2)!} = \boxed{e}$$

A cleaner way to do this utilizes the fact that for non-negative variables X,

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} P(X \geq i)$$

The probability Isabella made *at least* k throws is the probability that her first $k-1$ darts were sorted in order of closeness to the bullseye, which has probability $\frac{1}{(k-1)!}$. Thus the total expected number of darts thrown is

$$\sum_{k=1}^{\infty} \frac{1}{(k-1)!} = \boxed{e}$$

Problem. (Difficulty: Medium) You have two coins that you know are biased somehow, and you have a uniform prior on what their biases are. Suppose you flip both of them three times and the first one comes up heads twice, and the second one comes up heads once. What is the probability that the first coin is more biased towards heads than the second coin?

Solution. Let p and q denote the biases (towards heads) of the two coins respectively. Given the results of the flips and the uniform prior of the biases of the coins, the probability of coin 1 having bias p is proportional to $p^2(1-p)$ and the probability of coin 2 having bias q is proportional to $q(1-q)^2$. Thus the answer is

$$\frac{\int_0^1 \int_0^p p^2(1-p)q(1-q)^2 \, dqdp}{\int_0^1 \int_0^1 p^2(1-p)q(1-q)^2 \, dqdp}$$

which is just an annoying bash but easily doable.

Problem. (Difficulty: Medium, Medium) Sample uniformly from the unit disk, assuming

(a) You can sample uniformly from the interval $[0, 1]$

(b) You can sample from the Gaussian distribution $N(0, 1)$ and uniformly from the interval $[0, 1]$, but the latter only once.

Solution. For part (a), there are multiple approaches that work. One simple approach is to do rejection sampling; sample $x, y \in [0, 1]$ and if $x^2 + y^2 > 1$, re-sample. The problem with this is it does not extend well to higher dimensions; as dimension increases, the probability of having to re-sample goes to 1!

Instead, we treat a point on the disk as having two identifiers; a radius, and an angle. We can randomly sample the angle by sampling $\theta \in [0, 1]$ and then multiplying by 2π. Sampling the radius r is a little more tricky - a big trap is to just try to randomly sample $r \in [0, 1]$. This does not work!!! Notice that on a unit disk, there are more points with a larger radius than with a smaller radius (can be seen by drawing circles and calculating circumferences). Thus instead we want to sample r^2 uniformly at random from $[0, 1]$, and from there take the square root to recover the radius r. This works because since area is proportional to r^2, we are now sampling uniformly with respect to area, which is what we want. Then we obtain our uniformly sampled point $(r\cos(2\pi\theta), r\sin(2\pi\theta))$.

The key for part (b) is that if we sample x, y from $N(0, 1)$, then the point $\left(\frac{x}{\sqrt{x^2+y^2}}, \frac{y}{\sqrt{x^2+y^2}}\right)$ is uniformly distributed on the perimeter of the unit disk (convince yourself of this using the pdf of the normal distribution). Then we sample a radius r using the same trick as from part (a) and multiply the point by r to get our uniformly chosen point.

Problem. (Difficulty: Hard) What is the expected number of die rolls until you see a 6, given that every roll came up even.

Solution. This is a troll problem. The natural approach is to say that if every roll came up even, the possibilities were $2, 4, 6$ and so the probability of seeing a 6 was $1/3$. This would lead to an answer of 3. This however is very wrong. The possibilities are still $1, 2, 3, 4, 5, 6$ - in fact, knowing that every

roll came up even implies that there probably weren't very many rolls, since if there were lots of rolls probably we would have seen an odd number.

The simple way of solving this problem is to consider rolling a die until you see $1, 3, 5$, or 6. The expected length of a sequence of die rolls that ends with 6 matches the expected length of a sequence of die rolls that ends with $1, 3$, or 5 so to solve the problem we can actually take the expected length over all of these sequences. And by simple geometric probability, that is given by $\boxed{3/2}$ since the chance of seeing a $1, 3, 5$, or 6 is $2/3$.

A less clever way to solve the problem is to simply use Bayes' rule. Let X denote the number of rolls until we see 6 and let E be the event that every roll came up even before we see a 6. Then

$$P(X = k|E) = \frac{P(E|X = k)P(X = k)}{P(E)}$$

Now we know

$$P(E|X = k) = \left(\frac{2}{5}\right)^{k-1}$$

$$P(X = k) = \frac{1}{6}\left(\frac{5}{6}\right)^{k-1}$$

$$P(E) = \frac{1}{6}\sum_{k=1}^{\infty}\left(\frac{1}{3}\right)^{k-1} = \frac{1}{4}$$

Therefore

$$E[X] = \frac{2}{3}\sum_{k=1}^{\infty} k\left(\frac{1}{3}\right)^{k-1} = \frac{2}{3}\left(\sum_{k=1}^{\infty}\left(\frac{1}{3}\right)^{k-1}\right)^2 = \boxed{\frac{3}{2}}$$

Problem. (Difficulty: Hard, Hard)

(a) Place n points randomly on a circle. Find the probability all n are contained on the same semicircle.

(b) Choose $n-1$ points randomly on a line segment and break it into n pieces. What is the probability you can form a convex n-gon with those n pieces.

Solution. For part (a), consider all the uncountably infinite ways you can put n points on a circle. Group them into groups of size 2^n, where two configurations are in the same group if you can recover one from the other by replacing some points with the points diametrically opposite to them on the circle. Note that for each of these groups, exactly $2n$ of the configurations have points that lie on the same semi-circle. The way to see this is to take each of the $2n$ potential points in a configuration in a given group and note that there is exactly one configuration in that group that includes that point and where every other point is on the semicircle emanating clockwise from the original point. Thus the answer is $\frac{2n}{2^n} = \boxed{\frac{n}{2^{n-1}}}$.

For part (b), instead of picking $n-1$ points on a segment, realize this is equivalent to picking n points on a circle and using the arcs between them as side lengths (by "bending" the segment). Then by the Triangle Inequality, you can only be unable to form a polygon if all of the points lie on one semicircle (because then you'll have one really long side). Thus using part (a) the answer is $\boxed{1 - \frac{n}{2^{n-1}}}$.

Finance

Problem. (Difficulty: Easy) You have a strategy that you think has Sharpe ratio 8. Suppose that after n days of live trading, it has lost money. What is the value of n for which you reject your assumption that the strategy has a Sharpe of 8?

Solution. An annual Sharpe of 8 corresponds to a daily Sharpe of $8/\sqrt{252} \approx 8/16 = 0.5$ since there are 252 trading days in the year. Thus, assuming the strategy trades at a constant volatility, the z-score of being in the negative after n days is given by $0.5\sqrt{n}$. We reject the hypothesis when the z-score is greater than 2, or $\boxed{n = 16}$ days. Note that the rejection threshold of "2" was arbitrarily chosen, but any reasoning along these lines is correct.

Problem. (Difficulty: Easy) What is an advantage of using monthly return data over daily return data?

Solution. Monthly return data is easier to store (based on storage capacity) and also monthly returns look much more normally distributed than daily returns, which are more prone to have fat tails and extreme outliers.

Problem. (Difficulty: Medium) Suppose there are two derivative Contracts on a stock, currently trading at \$90, that work as follows. Contract A pays \$1 dollar if in one week, the stock's price ends above \$100. Otherwise, it pays \$0. Contract B pays \$1 if at any time during the next week the stock price crosses \$100, even if later in the week it falls below \$100. Otherwise, it pays \$0. If Contract A is trading for thirty cents, how much would you expect Contract B to trade for?

Solution. Assume that the stock's prices move randomly, and that its price is equally likely to go up or down. In every scenario where Contract B doesn't pay out, Contract A also doesn't pay out. In every scenario where Contract B does pay out, there is some point during the week where the stock price hits \$100 exactly. From that point consider the curve of the price until the end of the week. The probability of that curve occurring is equal to the probability that the horizontal mirror of that curve occurs. In one of these scenario, Contract A pays out, and in the other, it doesn't. This implies that every scenario in which Contract A pays out can be mapped to two equally likely scenarios in which Contract B pays out, so Contract B should trade at $\boxed{\text{sixty cents}}$, or twice what Contract A trades for.

Problem. (Difficulty: Hard) You have \$100 and are betting on a fair coin flip. You can bet any percentage of the \$100. If you win, you gain 1.2 times your bet (and your bet back), but if you lose, you lose your bet. What is the optimal bet size to maximize long-run expected earnings?

Solution. The way to answer these problems is known as the *Kelly Criterion.* Lots of places ask about the Kelly Criterion so you should definitely look it up - the derivation comes from maximizing log-utility (in this case, just the log of total wealth).

Problem. (Difficulty: Hard) What would a good model for market impact be?

Solution. Implicitly, this is asking for a model of the book. Intuitively, there will be only a few people willing to trade at the best bid/offer, more people willing to trade at the one tick below/above the best bid/offer, etc... Thus the order book probably looks like a pyramid. Thus to trade $\$x$ worth of an asset by just hitting bids or lifting offers off the book, you'd expect the average price you get to be $\Theta(\sqrt{x})$ away from the midpoint price. Therefore total market impact as a function f of the size of your trade x (in dollars) can be approximated as $\boxed{f(x) = cx^{3/2}}$ for some constant c depending on things like the average daily volume of the stock, its volatility, etc...

Problem. (Difficulty: Hard) Suppose you have a strategy that produces a daily stream of mean-0 forecasts on a stock $f_1, f_2, \ldots$ and that the daily mean-0 returns of that stock are given by $r_1, r_2, \ldots$. Suppose that the way your strategy works is that on every day i your dollar position in the stock is

equal to f_i, so that on day i your return is $f_i r_i$. Suppose that the daily Sharpe of your strategy is 0.05, and that your forecasts have 0 auto-correlation. Suppose the forecast on day i is uncorrelated to every return not on day i, so that $\mathbb{E}[f_i r_j] = 0$ for all $i \neq j$. Finally, suppose that your forecasts and the returns have variance 1. Find a way to transform your forecasts so that they have auto-correlation $\rho > 0$, and compute the expected effect on the Sharpe of your strategy.

Solution. Since the Sharpe of the strategy is 1, we have that $\mathbb{E}[f_i r_i] = 0.05$. Now suppose we transform the forecasts f into forecasts g using an exponentially-weighted moving sum as follows:

$$g_i = f_i + \rho f_{i-1} + \rho^2 f_{i-2} + \dots$$

Then we have that

$$\mathbb{E}[g_i g_{i-1}] = \mathbb{E}\left[(f_i + \rho g_{i-1}) g_{i-1}\right] = \rho \mathbb{E}[g_{i-1}^2]$$

so the auto-correlation of g is ρ as desired. Also note that

$$\mathbb{E}[g_i r_i] = \mathbb{E}[f_i r_i] = 0.05$$

since the forecasts on days before i are uncorrelated to the return on day i. Finally, note that

$$\mathbb{E}[g_i^2] = 1^2 + \rho^2 + \rho^4 + \dots = \frac{1}{1-\rho^2}$$

This implies that the daily Sharpe of the new strategy is $\dfrac{\mathbb{E}[g_i r_i]}{\sqrt{\mathbb{E}\left[(g_i r_i)^2\right] - \mathbb{E}\left[g_i r_i\right]^2}} \approx \dfrac{\mathbb{E}[g_i r_i]}{\sqrt{\mathbb{E}\left[g_i^2\right]}} = \boxed{0.05\sqrt{1-\rho^2}}$. Unsurprisingly, slowing our strategy down leads to a small reduction in Sharpe.

Data Science

Problem. (Difficulty: Easy) If we are using a simple linear regression and we add a feature, what will happen to R^2 on the training set?

Solution. It will either stay the same or increase. The easiest way to see this is that in the worst case, the coefficient on the new feature in the linear regression is 0, which would make the R^2 stay the same.

Problem. (Difficulty: Easy) How to prevent overfitting?

Solution. There are a ton of answers you can give to this; I would give as many as possible.

- Have a strong, fundamental, testable hypothesis before performing an experiment and try to use Bayes' rule to combine the strength of your prior with the results of the experiment.
- Start with the simplest possible experiments, using the fewest number of parameters possible.
- Have a clear train/test split, decided upon before looking at any data. Some notes about this: low signal-to-noise data like technical financial data often benefits from having a similarly-sized train and test sets.
- Use regularization - e.g. ridge regression/lasso regression versus simple linear regression. Note that most regularization methods that adjust a loss function re actually grounded in Bayesian priors on your parameters.
- Use dimensionality reduction (e.g. PCA, etc...) to limit the number of features you might overfit on.
- Use early stopping during training, if applicable.

- Use robust evaluation techniques. For example, k-fold cross-validation, AIC or BIC, adjusted R^2, etc...

Problem. (Difficulty: Easy) Suppose you have a data set with a large number of features. How will you handle the large number of features?

Solution. There are a ton of answers you can give it to this; I would give as many as possible.

- Use fundamental, hypothesis-driven reasoning to select a few features that you think are predictive, and begin by working with those
- Use feature-selection algorithms like stepwise linear regression and lasso regression to determine which features are predictive
- Use non-linear models like random forest that let you easily look at feature importances, to determine which features are more likely to be predictive
- Use dimensionality reduction (e.g. PCA) to combine your features into a more reasonable number of synthetic features.

Problem. (Difficulty: Easy) How to deal with categorical variables (e.g. country of origin)?

Solution. Different models per value of the variable, or one-hot vectors.

Problem. (Difficulty: Easy) Describe the bias-variance tradeoff.

Solution. https://en.wikipedia.org/wiki/Bias-variance_tradeoff

Problem. (Difficulty: Medium) Suppose I ran a linear regression on a dataset, and obtained a β, a t-stat, and an R^2. Now suppose I accidentally duplicated my dataset and re-ran the regression - how would these three metrics change?

Solution. β and R^2 would stay the same, t-stat would increase by a factor of $\sqrt{2}$.

Problem (Difficulty: Medium, Medium, Hard)

(a) Derive the closed-form solution to simple linear regression

(b) Explain the advantages and disadvantages of using this closed form versus an approximation found from SGD

(c) Explain ridge regression from a Bayesian perspective and derive a closed-form solution

Solution.

(a) Suppose our response values are stored in an $n \times 1$ matrix Y and we have n data points concatenated to an $n \times m$ matrix X, where each data point has m features. Then the goal is to find an $m \times 1$ matrix β so that βX approximates Y. Specifically we want to minimize $(Y - X\beta)^T(Y - X\beta)$. Differentiating this expression with respect to β yields $X^T(Y - X\beta)$ and so setting this equal to 0 we find that the solution is $\boxed{\beta = \left(X^T X\right)^{-1} X^T Y}$

(b) The advantage over SGD is that it is exact; the disadvantage is that it requires matrix multiplication and inversion which is slow, while SGD is much faster.

(c) Frame the linear regression problem in terms of probability: say that the log-likelihood of observing data X, Y is equal to

$$e^{-(Y-X\beta)^T(Y-X\beta)}$$

Note that given this form there is an implicit assumption that the error terms in this regression are drawn from a homoscedastic normal distribution. If we just wanted to find the β that maximized log-likelihood, this would be equivalent to the original linear regression setup. But, suppose we had a prior that each element in β should be close to 0. More precisely suppose that our prior was $\beta_i \sim N(0, \sigma)$ for all i and some $\sigma \in \mathbb{R}$. Then the log-likelihood would become

$$e^{-(Y-X\beta)^T(Y-X\beta)} \cdot e^{-\frac{\beta^T\beta}{2\sigma}}$$

where the second term represents the prior. Taking the log of this expression then yields exactly the expression to minimize for ridge regression, namely

$$(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$$

where $\lambda = (2\sigma)^{-1}$. Solving this minimization problem the same way as part (a), we get a solution of $\boxed{\beta = \left(X^TX + \lambda I\right)^{-1} X^TY}$

Computer Science

Problem. (Difficulty: Easy) Efficiently find an integer that uses each of the digits $1, 2, \ldots, 9$ exactly once that has the number determined by its first k digits divisible by k for all k.

Solution. The most natural approach is to implement DFS. A much cheekier, quick solution is to notice that there are less than $400,000$ permutations of those digits and you can just try permutations at random. You expect to find the right number after around $200,000$ random attempts, which the computer can do very quickly. Less cheeky but also quick is to try every possible permutation, which again works well due to the small search space.

Problem. (Difficulty: Easy, Hard)

(a) Write code to calculate the nth Fibonacci number in $O(n)$ time

(b) Do it in $O(\log n)$ time.

Solution. For part (a), just initialize a list L of length $n+1$ with the first two elements being 0 an 1 and for i in range 2 to n set $L[i] = L[i-1] + L[i-2]$, and return $L[n]$.

For part (b), notice that we have the matrix equality

$$\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}^n = \begin{pmatrix} F_{n+1} & F_n \\ F_n & F_{n-1} \end{pmatrix}$$

for all $n \geq 1$. Thus we can compute the LHS using exponentiation by squaring in $O(\log n)$ time and recover the nth Fibonacci number that way. You can do this with Binet's formula as well, but keeping track of necessary significant digits is annoying.

Problem. (Difficulty: Medium) Write a recursive function to calculate the number of partitions of a positive integer.

Solution. Let $p(n, m)$ denote the number of partitions of n into parts all no larger than m. Then we can split these partitions into two groups: those that contain m in the sum and those that don't. This leads to the natural recurrence:

$$p(n, m) = p(n, m-1) + p(n-m, m)$$

from which is is easy to calculate $p(n, n)$, the total number of partitions of n, efficiently.

Problem. (Difficulty: Medium, Medium, Hard)

(a) Suppose you have an n-story building and two eggs, and you want to find out which is the highest floor from which a dropped egg will not crack. If you drop an egg and it doesn't crack, you can re-use it in your experiment. What is the minimum number of drops you need to attempt to discover the answer?

(b) Using dynamic programming generalize to k eggs, and give a time-complexity for your algorithm.

(c) Do better than part (b)

Solution. We solve both (a) and (b) at once. Let $f(n, k)$ denote the minimum number of drops for an n-story building with k eggs. Note that $f(n, 1) = n$ since the worst-case scenario is that it doesn't crack if dropped from any floor, and you would have to test every floor sarting from the ground floor one at a time.

Now suppose you have $k > 1$ eggs and you start by dropping the first egg from floor m. Then if it cracks, you have reduced the problem to the problem with an $m - 1$ story building with $k - 1$ eggs. Otherwise, you have reduced the problem to an $n - m$ story building with k eggs, since you only need to consider floors above floor m. Thus we have the recurrence:

$$f(n, k) = 1 + \min_m \max\left(f(m - 1, k - 1), f(n - m, k)\right)$$

which allows you to find the value of $f(n, k)$ in $O(n^2 k)$ time.

For part (c), it turns out that you can find a closed form for the largest possible building height you can solve the problem for if you have k eggs and are allowed d drops. The closed form turns out to be

$$\sum_{i=0}^{k} \binom{d}{i}$$

which you can prove with recursion, and then the original problem amounts to finding the smallest value of d for which

$$\sum_{i=0}^{k} \binom{d}{i} \geq n$$

A simple binary search then takes $O(k \log n)$ time, a massive improvement over the $O(n^2 k)$ time from part (b).