Additional midterm practice problems for CS109

Disclaimer: You should prioritize official CS109 review materials (past exams, homeworks, sections) first! These problems are not guaranteed to match actual CS109 exam problems in style, scope, or difficulty (mine probably skew harder, and some of them involve calculations that would be too onerous by hand). Hence "midterm practice problems" rather than "practice midterm problems". These are unofficial and were written for Ian's CS109A class for Winter 2022; anyone is welcome to share them around, but please direct any questions about the problems to `itullis@stanford.edu`. Star ratings: *, ** = bread and butter, *** = extra thought required, **** = stretch

1 Revenge of the past problems

(a) (*) This problem is based on Section 4, Problem 2, but you should not look at that problem while solving this one! Assume that people who are suffering from a certain illness (which is not COVID because goodness knows we've had to think about that enough) have a temperature distributed as $\mathcal{N}(\mu = 101, \sigma = 1)$. However, people who don't have the illness have a temperature distributed as $\mathcal{N}(\mu = 98, \sigma = 1)$.

Suppose that we observe a person with a temperature of 101 or higher. Why shouldn't we necessarily conclude that the person has the illness, even though 101 clearly has a larger probability density in the "illness" distribution than in the "non-illness" distribtion? (I.e., what other piece of information do we need before making our decision? The intended answer is *not* that there are other diseases in the world too, although that is a reasonable point... but here, suppose that we have somehow narrowed it down to the person either having that illness or having no illness.)

(b) (**) In Homework 3, Problem 3, when testing for measles (in a population with a 5% rate of measles), we tested pooled samples of 6 people at a time, then tested them all individually if there was a positive pooled test. Suppose that we replaced 6 with N. What is the smallest value of N (greater than 1, of course) for which this strategy becomes *worse* (in terms of expected total number of tests) than just testing everyone individually at the outset? (Use e.g. Wolfram Alpha for calculations.)

(c) (***) In Homework 3, Problem 1, suppose that the question had used a **fair** coin, and 11 flips. Can you see how to immediately answer the question "what is the probability that the number of heads is odd?" without doing any math? (Optional **** extension: what if it were an even number of flips, like 12?)

(d) (**) On the fall 2021 midterm, Problem 3ab, we observed a 0 degree (non)-movement and then a 16 degree movement. Suppose that we had started with the same prior of $\frac{3}{4}$ but made those observations in the opposite order, i.e., we saw 16, then updated our estimate and used that as the new prior, and then saw 0, and updated our estimate again. Would you expect the final estimated probability to be different? Is it?

2 2-parameter distribution, 1-parameter approximation?

(*) When we use a normal distribution to approximate a binomial distribution, we take the mean μ_B and variance σ_B^2 of the binomial, and then use those directly as the mean μ_N and variance σ_N^2 of the normal. (These subscripted variables are not something we've seen in class; the subscripts are just there for clarity.)

But when we use a Poisson approximation, it's a bit more awkward since the Poisson distribution only has the single parameter λ as both its mean and variance. So we have to hope that both μ_B and σ_B^2 are close to λ, which also implies that we need μ_B and σ_B^2 to be very close to each other. When (in terms of the n and/or p parameters of the binomial distribution) would you expect this to be true? Does this fit your intuition about when it's OK to use a Poisson approximation to a binomial?

3 Central "Express"way, except when it isn't!

As usual, I am running late, so I leave home 25 minutes before our CS109A class starts. Suppose that there are 16 traffic lights on my drive from home to campus. Also suppose that if I hit only green lights, it would take me 20 minutes to reach campus. However:

- Each light has (independently) a probability of 0.3 of being red.
- For each light (independently), the amount of additional time I have to wait at that light is normally distributed with mean 60 seconds and standard deviation 15 seconds.

(a) (*) What is the expected amount of time it will take me to reach campus?

(b) (***) What is the **exact** probability that I make it to campus in time to teach class? (Assume, somewhat unrealistically, that I teleport instantaneously from my parking spot on campus to our classroom, which is obviously not true since it's up a ton of stairs.) Your answer can involve a summation and Φ.

You can use the fact that the sum of n independent $\mathcal{N}(\mu, \sigma^2)$ is itself a normal distribution $\mathcal{N}(n\mu, n\sigma^2)$; we haven't quite gotten there in lecture yet. So if I hit 3 lights, for instance, my total waiting time is distributed as $\mathcal{N}(3 \cdot 60, 3 \cdot 15^2)$. However, this question *does* test something that could be on the midterm! What distribution can you use for the number of lights I hit?

(c) (***) Suppose that the traffic lights are not independent, in some way that guarantees that I will never hit two red lights in a row, but that the probability of hitting each one is still 0.3 overall. (As one example, suppose that with probability 0.5, each odd-numbered light is red with probability 0.6 and even-numbered lights are never red, and with probability 0.5, each even-numbered light is red with probability 0.6 and odd-numbered lights are never red.) Would the answer to (a) to be larger, smaller, or the same, or would it depend on the specific way in which these conditions hold?

4 World Series of Baseball

The World Series is only 7 games, which is not enough to decide which of two teams of similar ability is actually better. But the format is presumably designed to create drama and excitement, not to select the worthiest champion!

(**) Suppose that in any given World Series game, the Dodgers have a true probability of 0.51 of beating the Giants[1], independently of the results of any other games. What is the minimum *odd* number of games that the World Series would need to have in order for there to be less than a 1% chance of the Giants winning a majority of the games? (The number of games must be odd to avoid an overall tie.) **Use a normal approximation.** You may (and should) use the fact that $\Phi(2.3263) = 0.99$.

5 Hit point insecurity

In Dungeons and Dragons, I always dread rolling dice to determine my character's hit points. What if I get unlucky and roll a 1? What if someone else gets so lucky that their sneaky rogue has as many hit points as my beefy fighter?

Suppose that I determine my fighter character's hit points (H_F) by rolling ten 10-sided dice and adding them together, and my friend determines her rogue character's hit points (H_R) by rolling ten 6-sided dice and adding them together.

(a) (*) What is $\mathbb{E}(H_F)$?

(b) (**) What is $Var(H_F)$?

(c) (****) What is the **exact** probability that my friend's rogue has **exactly** the expected number of hit points for a fighter, i.e., $P(H_R = \mathbb{E}(H_F))$? Your answer must be a single term with a single $\binom{n}{k}$ type expression in the numerator.

(d) (***) Suppose that my Dungeon Master allows me to reroll any 1s that come up when I roll my ten 10-sided dice, but only once each (i.e. even if a rerolled 1 comes up 1 again, I have to keep that 1). What is $\mathbb{E}(H_F)$ in this case?

(e) (***) Suppose that my Dungeon Master allows me to reroll any 1s that come up when I roll my ten 10-sided dice, and then keep rerolling any 1s, and so on, until there are no more 1s. What is $\mathbb{E}(H_F)$ in this case?

[1] Apologies to Giants fans.

6 Not quite six sigma

(a) (*) In a normally distributed population, what fraction of the population do you equal or exceed if you are 1 standard deviation (σ) above the mean? 2σ? deviations? 3σ? (This wouldn't be a midterm question, because it can't be hand-solved, but: do you know how to find this info? I do think it's useful (for life) to memorize these numbers...)

(b) (*) Tests like the SAT have often been scaled based on an assumed normal distribution with a mean of 500 and a standard deviation of 100. However, scores are reported rounded to the nearest 10-point increment. What fraction of test-takers would you expect to earn a score of 670? (This isn't a trick question where the answer is 0... what range of the distribution does this actually represent?)[2] You can leave your answer in terms of Φ expressions.

7 Critical success and critical failure

(a) (**) Suppose you roll a 20-sided die until you see a 20. What is the expected number of rolls that this will take?

(b) (**) Suppose you roll 20 20-sided dice. What is the probability that you will see at least one 20?

(c) (***) Suppose you roll a 20-sided die until you have seen *both* at least one 20 *and* at least one 1. What is the expected number of rolls that this will take? (Hint: consider breaking the process down into two phases.)

(d) (****) Suppose you roll 20 20-sided dice. What is the probability that you will see at least one 20 and at least one 1? (This is hard! Feel free to leave your answer in terms of summations etc. Also consider breaking the problem down into cases that you know are mutually exclusive and exhaustive, so that you don't have to worry about trying to correct for double-counting.)

8 An easy mistake

(**) When we roll 5 6-sided dice, the probability of seeing two of one number and three of a different number – e.g., 52522 – is $\frac{6\cdot 5\cdot \binom{5}{2}}{6^5}$. Why is it that when we roll 4 6-sided dice, the probability of seeing two of one number and two of a different number is **not** $\frac{6\cdot 5\cdot \binom{4}{2}}{6^4}$?

BEST OF LUCK ON THE MT! Solutions to these problems begin on the next page...

[2] This question should not be taken as an endorsement of standardized testing, which has its problems, to say the least. Nor do one's test scores determine one's worth in any way. That said, you should still try to do your best on the midterm!

1. (a) The issue is that we do not know how common the illness is overall. Suppose that the illness is extremely rare, occurring in, e.g., 0.1% of the population. But we would expect about 0.13% of the non-illness population to have this temperature or higher! (This comes from evaluating the normal CDF with $\mu = 98$, $\sigma = 1$, $x = 101$, and then subtracting the result from 1 because we want the area to the right of that point, under the long tail of the curve.) So in this case, the fraction of the population with the illness is pretty similar to the fraction of the population that doesn't have the illness but just happens to have that high of a temperature. Therefore, it's hard to say which of those two groups the person is in!

   The takeaway point is that we need to know **the overall frequency of the illness** to make a sensible prediction. Notice that the section problem had to give you this piece of information (it's 20% in that case). A similar situation arose in that problem, where even though 100 "looks" much more like a temperature from the group with the flu, the proportion of people with the flu also factors into the calculation, and so the answer ends up actually being not far off from 50%.

   (b) On the homework, we found that the expected number of tests was $1(0.95^6)+(1+6)(1-0.95^6)$. Here, we replace 6 with the more general N: $1(0.95^N)+(1+N)(1-0.95^N)$. The pooled-test strategy becomes worse once that quantity exceeds N (which is the number of tests we would need if we skipped the pooled test and just tested everyone up front). So, to get the threshold N above which the pooled-test strategy is worse, we solve

   $$1(0.95^N) + (1 + N)(1 - 0.95^N) = N$$

   $$0.95^N + 1 - 0.95^N + N - N \cdot 0.95^N = N$$

   $$1 = N \cdot 0.95^N$$

   Using Wolfram Alpha, we find that $N \approx 1.05$ or $N \approx 87.08$. Both of these are thresholds, but only one is the kind we want (where the pooled test goes from useful to not useful).

   Let's check for $N = 87$: $1(0.95^{87}) + (1 + 87)(1 - 0.95^{87}) \approx 86.997 < 87$ - still good!

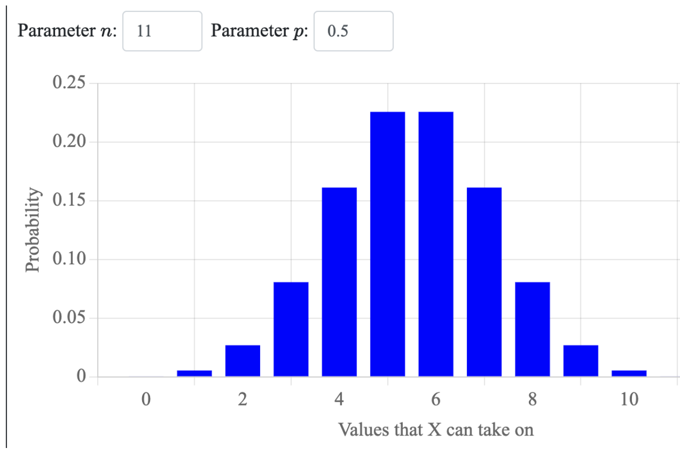
   And for $N = 88$: $1(0.95^{88}) + (1 + 88)(1 - 0.95^{88}) \approx 88.036 > 88$ - no longer good!

   So the answer is $\boxed{88}$.

   (c) A binomial distribution with $p = 0.5$ is always symmetric. To see why, consider that when $p = 0.5$, regardless of the value of n, $P(X = x) = \binom{n}{x}0.5^x(1-0.5)^{n-x} =$

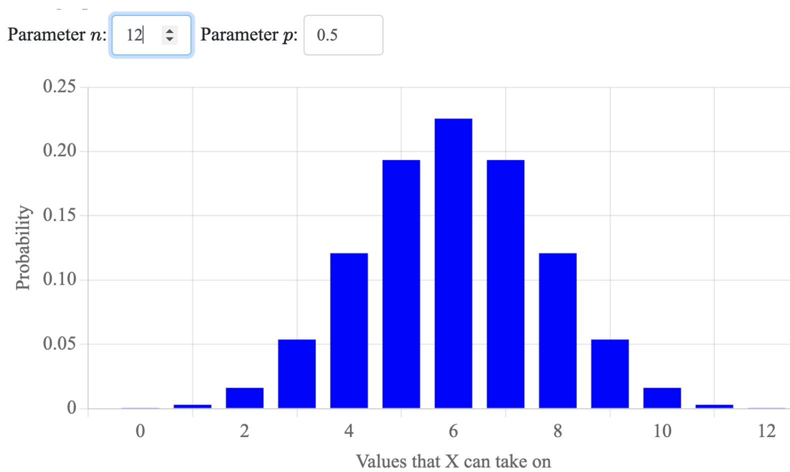
$\binom{n}{x}0.5^{x+n-x} = \binom{n}{x}0.5^n$, and $P(X = n-x) = \binom{n}{n-x}0.5^{n-x}(1-0.5)^{n-(n-x)} = \binom{n}{n-x}0.5^{n-x+x} = \binom{n}{n-x}0.5^n$. Since as a rule, $\binom{n}{x} = \binom{n}{n-x}$, these quantities are the same. (As a reminder about why that last fact is true, picking x out of n things is the same as choosing which $n-x$ of n things to leave behind.)

Here's the distribution for $n = 11, p = 0.5$. (The values for $x = 0$ and $x = 11$ are nonzero, but too small to show up here.):



If we look at the bars for the odd values of x (1, 3, 5, 7, 9, 11), they are the same as the bars for the even values of x, looking the other way (10, 8, 6, 4, 2, 0). So the two sets of bars must have the same sum, specifically, $\boxed{\frac{1}{2}}$.

What about for $n = 12$?



Now we have a problem: we can no longer make the same symmetry argument. For instance, the bar for $x = 6$ does not have a counterpart in the odd numbers. Yet, probably surprisingly, it is still true that the answer is $\frac{1}{2}$.

Here's an intuitive argument about why this is true. Consider making 12 coin flips and keeping track of just the parity (odd or even) of how many heads we have seen. We start with 0 heads, which is even parity. Every time we make a flip, we have a 50% chance of changing our parity and a 50% chance of leaving it the same. So after 1 flip, we have a 50% chance of having an even parity and a 50% chance of having an odd parity. After 2 flips, we still have a 50% chance of an even parity: 25% of the time, we were at even and we stayed there, and 25% of the time, we were at odd but then changed parity to even. This logic holds throughout the entire sequence of flips, regardless of whether n is odd or even.

(d) I am hoping that you expected the answer to stay the same. It would look really bad for Bayesian statistics if seeing the observations in a different order somehow changed our overall conclusion! But we should check:

Initial estimate: $P(\text{can hear sound}) = \frac{3}{4}$

After observing 16:

$P(\text{can hear sound}|\text{observe 16}) = \frac{P(\text{observe 16}|\text{can hear sound}P(\text{can hear sound}}{P(\text{observe 16})}$

$=\frac{P(\text{observe 16}|\text{can hear sound})P(\text{can hear sound})}{P(\text{observe 16}|\text{can hear sound})P(\text{can hear sound})+P(\text{observe 16}|\text{can't hear sound})P(\text{can't hear sound})}$

$= \frac{0.20\cdot\frac{3}{4}}{0.20\cdot\frac{3}{4}+0.08\cdot\frac{1}{4}} = \frac{0.15}{0.17} = \frac{15}{17}.$

Then, after observing 0, we use our new estimate of $\frac{15}{17}$ for the prior probability $P(\text{can hear sound})$.

$P(\text{can hear sound}|\text{observe 0})$

$= \frac{P(\text{observe 0}|\text{can hear sound})P(\text{can hear sound})}{P(\text{observe 0}|\text{can hear sound})P(\text{can hear sound})+P(\text{observe 0}|\text{can't hear sound})P(\text{can't hear sound})}$

$= \frac{0.08\cdot\frac{15}{17}}{0.08\cdot\frac{15}{17}+0.40\cdot\frac{2}{17}} = \frac{\frac{6}{85}}{\frac{10}{85}} = \boxed{\frac{3}{5}}.$

So even though we were very confident ($\frac{15}{17}$) that the baby could hear the sound after the first observation, the second observation drops that probability to $\frac{3}{5}$, the same as the original answer.

2. The mean and variance of a binomial distribution with parameters n and p are np and $np(1-p)$, respectively. (A good way to be able to rederive the variance is to think of a binomial distribution as a sum of n Bernoullis, and note that the variance of an individual Bernoulli trial is $p(1-p)$.)

If we want $np \approx np(1-p)$, then we see that we want $1-p \approx 1$, i.e. p is very close to 0. This makes sense – the Poisson distribution is a binomial taken to the limit in which the number of trials is very large and the success probability is very small. This also matches the rule of thumb that Poisson approximations work well when p is small.

If we try to use a Poisson approximation when e.g. $p = 0.5$, we will probably find that it does not work very well. For instance, if we try to use a Poisson to model the number of heads in 100 fair coin flips given that the expected number is 50, and we ask about $P(X = 40)$, we get $\frac{e^{-50}50^{40}}{40!}$, which is ≈ 0.021. But the actual answer is $\binom{100}{40}(0.5)^{40}(1-0.5)^{60} \approx 0.011$. So the Poisson approximation is way off!

3. (a) Each traffic light has an 0.3 probability of adding an expected 60 seconds of delay, so it adds 18 seconds in expectation. By linearity of expectation, the sum for all 16 lights is 288 seconds = 4 minutes, 48 seconds. So I will on average take $\boxed{24:48}$ to reach campus, which is cutting it close!

   (b) Suppose that I hit n lights. Then the total delay is the sum of n independent and identically distributed $\mathcal{N}(\mu = 60, \sigma^2 = 225)$ random variables. Per the rules for adding independent normal distributions, as given in the problem, the total delay is $\mathcal{N}(\mu = 60n, \sigma^2 = 225n)$. Then the probability that I make it to campus on time (i.e. that the delays are 300 seconds or less) is $\Phi(\frac{300-60n}{\sqrt{15n}})$. Since this delay time is continuous, there is no need for a continuity correction.

   However, the number of lights I hit is binomially distributed: $P(N = n) = \binom{16}{n}(0.3)^n(0.7)^{16-n}$. The distribution of the total delay depends on this value, so it is:

$$\boxed{0.7^{16} + \sum_{n=1}^{16}\binom{16}{n}(0.3)^n(0.7)^{16-n}\Phi(\frac{300-60n}{\sqrt{15n}})}.$$

   The separate 0.7^{16} term is there because we can't have the summation start at 0, or we would have a division by 0 in the Φ term. In that case, the Φ term wouldn't even make sense, because there would be no distributions involved! If I hit no red lights, then I always make it on time.

   Observe that earlier in the summation (for small values of n), the Φ part will be very close to 1 (since there are so few lights that they can't possibly cause enough of a delay), and later in the summation, the Φ part will be very close to 0 (since there are so many lights that I can't possibly avoid being too delayed) *and* the binomial part will be close to 0 (since it is unlikely to see so many red lights, given that they are individually uncommon). If we do the math, it turns out that within the summation, only the $n = 1$ through $n = 4$ terms really matter.

   (c) Linearity of expectation holds regardless of the independence or nonindependence of the individual traffic lights; notice that the argument in (a) never invoked independence. So the answer is exactly the same regardless of the exact way the non-independence manifests.

4. Let W be a random variable representing the number of wins the Giants (not the Dodgers) get in n games, for some value of n that we have yet to find. The true distribution of W is binomial: $P(W = w) = \binom{n}{w}(\frac{49}{100})^w(\frac{51}{100})^{n-w}$. Per the usual formulas, the mean and variance are $np = \frac{49n}{100}$ and $np(1-p) = \frac{2499n}{10000}$, so we will use those as μ and σ^2 in our normal approximation.

   Then the probability that the Giants win over half the games is

   $P(W > \frac{n}{2}) = 1 - \Phi(\frac{\frac{n}{2} - \frac{49n}{100}}{\sqrt{\frac{2499n}{10000}}})$. Note the absence of a continuity correction, even though the number of wins is a discrete value (and there are also no ties in the World Series! See Rule 7.02(a) Comment in the MLB Official Baseball Rules.) This is because we are specifically assuming that n is odd. Suppose, for example, that $n = 7$; then values of 3.4 and 3.6 in continuous-land would already fall into the 3 and 4 buckets, respectively, in discrete-land, which is exactly what we want.

   For the above probability to be 0.01, we need $\Phi(\frac{\frac{n}{2} - \frac{49n}{100}}{\sqrt{\frac{2499n}{10000}}}) = 0.99$. Using the provided piece of information about Φ, we know that $\frac{\frac{n}{2} - \frac{49n}{100}}{\sqrt{\frac{2499n}{10000}}} = 2.3263$. Solving this for n (which involves squaring to get rid of the $\sqrt{n}$ and then using the quadratic equation – again, you would probably not have to do this by hand on an exam, and I just used Wolfram Alpha), we find that $n \approx 13523.8$. Because we are requiring n to be odd, the next largest odd value is $\boxed{13525}$. Since Ian eats at least one chocolate malt per baseball game, that's a lot of chocolate malts, especially if the Series makes it all the way to that crucial game 13525...

   Let's double-check that that makes sense. The mean is 6627.5 and the variance is 3379.8975. The probability that the Giants win more than half the games in the series – i.e. 6762.5 – is $1 - \Phi(\frac{6762.5 - 6627.5}{\sqrt{3379.8975}}) = 1 - \Phi(2.322) = 0.01$.

   What is the real answer without the approximation? We want the smallest odd n such that $P(W_{\text{Giants}} > \lceil \frac{n}{2} \rceil) = \sum_{w=\lceil \frac{n}{2} \rceil}^{n} \binom{n}{w}(\frac{49}{100})^w(\frac{51}{100})^{n-w} < 0.01$. This was extremely painful for my computer to check in Python (check out the `Decimal` library if you need to work with gigantic factorials, and consider using PyPy), but it found the answer 13527. So the normal approximation was very close! (and the error probably came from how I provided 2.3263 in a truncated form, not from the failure of the Central Limit Theorem)

5. (a) For a single roll R, the expected value $\mathbb{E}(R)$ is $\frac{1+2+...+10}{10} = \frac{11}{2}$. So $\mathbb{E}(H_F)$ is 10 times that, i.e., $\boxed{55}$.

(b) We can use $Var(R) = \mathbb{E}(R - \mathbb{E}(R))^2$. The differences between 1, 2, ..., 10 and the mean are $-\frac{9}{2}, -\frac{7}{2}, -\frac{5}{2}, -\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \frac{7}{2}, \frac{9}{2}$. So $\mathbb{E}(H_F - \mathbb{E}(H_F))^2 = \frac{(-\frac{9}{2})^2 + (-\frac{7}{2})^2 + \ldots + (\frac{9}{2})^2}{10} = \frac{33}{4}$. Then (because the rolls are i.i.d) the variance is 10 times that: $\boxed{\frac{330}{4}} = 82.5$.

Alternatively, for our single roll R, we could use $Var(R) = \mathbb{E}(R^2) - \mathbb{E}(R)^2$. The first term is $\frac{1+4+\ldots+81+100}{10} = \frac{77}{2}$. Then $Var(R)$ is $\frac{77}{2} - (\frac{11}{2})^2 = \frac{33}{4}$ as before.

(c) This is tricky, but we can observe that in order for the rogue to get exactly 55 out of the possible 60 hit points, my friend must roll something close to all sixes, but with exactly 5 one-point deductions distributed among the ten dice. We can safely use the divider method here because even if all five deductions end up on the same die, this is still a valid result (a roll of 1). So we need to distribute 5 things among 10 buckets, and per the usual divider method formula, there are $\binom{5+10-1}{10-1} = \binom{14}{9}$ ways to do so.

This gives us the event space, and the sample space is all ways of rolling the 10 dice: 10^{10}. (Notice that in both the numerator and the denominator, the order of the rolls matters.) So the overall answer is $\boxed{\frac{\binom{14}{9}}{10^{10}}}$, which is very small (about 1 in 5 million).

Remark: This would not have worked with 54, for instance, since then the divider method could potentially assign all six deductions to the same die, which is not possible.

(d) As usual, let's start by thinking about just one die roll. $\frac{9}{10}$ of the time, we keep the original value on the die, and the other $\frac{1}{10}$ of the time, it is like a single new standard roll. Therefore the expected value of the single die is $\frac{1}{10}(2) + \ldots + \frac{1}{10}(10) + \frac{1}{10}(\mathbb{E}(R)) = \frac{54}{10} + \frac{1}{10}(\frac{11}{2}) = \frac{119}{20}$. Then $\mathbb{E}(H_F) = \boxed{\frac{119}{2}} = 59.5$. This is a noticeable improvement over the mean of 55 without the DM's generosity.

(e) Now when we get a 1, it is like we are beginning the entire rolling process over for that die, so we can write a recursive expression for a single die: $\mathbb{E}(R) = \frac{54}{10} + \frac{1}{10}\mathbb{E}(R)$. Solving for $\mathbb{E}(R)$, we find that it equals 6, so $\mathbb{E}(H_F) = \boxed{60}$. Allowing those infinite rerolls doesn't get us much more than the single reroll in part (d); intuitively, this is because multiple rerolls are so rare.

6. (a) We can get these values from the course reader's Gaussian CDF calculator: they are $\Phi(1), \Phi(2), \Phi(3)$, which are $\boxed{0.8413, 0.9772, 0.9987}$.

(b) 670 would correspond to anything between 665 and 675; anything just lower than 665 would be rounded to 660, and anything just higher than 675 would be rounded to 680. So we want the integral from 665 to 675 of a normal distribution with mean 500 and standard deviation 100. As usual, we don't actually do this integral, but instead frame it as the difference between two evaluations of the CDF: $\boxed{\Phi(\frac{675-500}{100}) - \Phi(\frac{665-500}{100})}$, which is $\approx 0.9599 - 0.9505 \approx 0.0094$.

7. (a) Because we want the total number of trials until one success, this is a geometric distribution with $p = \frac{1}{20}$. The expectation for a geometric distribution is $\frac{1}{p}$, so here it is $\frac{1}{\frac{1}{20}} = \boxed{20}$.

   (b) The probability of seeing at least one 20 is 1 minus the probability of seeing no 20s, i.e., every roll comes up something other than 20. That probability is $(\frac{19}{20})^{20}$, so the answer is $\boxed{1 - (\frac{19}{20})^{20}}$.

   We implicitly used a binomial distribution there, and we would get the same answer by using one directly, with $n = 20$, and $p = \frac{1}{20}$ being the probability of rolling a 20 on any given roll: $1 - P(X = 0) = 1 - \binom{20}{0}(\frac{1}{20})^0(\frac{19}{20})^{20}$.

   (c) Now we can't use a geometric or negative binomial distribution directly, even though the situation is in the same ballpark. Per the hint, let's think about two phases: the first phase, when we are trying to see our first 1 or 20 (it doesn't matter which we encounter first), and the second phase, when we are trying to see our first of the two numbers that we don't already have.

   The first phase is modeled by a geometric distribution, but with a success probability of $p = \frac{2}{20} = \frac{1}{10}$, since either a 1 or a 20 ends this phase. The expected number of rolls for this phase is $\frac{1}{\frac{1}{10}} = 10$.

   The second phase is also a geometric distribution, but now the success probability is down to $\frac{1}{20}$, since there is only one number we still need. (During this time, we might see additional copies of the number we already have.) The expectation here is 20 rounds, as before.

   Therefore the overall expectation is $10 + 20 = \boxed{30}$. We can add these because of linearity of expectation, and also because the two phases have been set up to be clearly non-overlapping.

   (d) This is hard to solve directly, but we can break it up into four mutually exclusive and exhaustive cases:

      i. There are no 1s or 20s.
      ii. There is at least one 1, but there are no 20s.
      iii. There is at least one 20, but there are no 1s.
      iv. There is at least one 1 and at least one 20.

Case i. is similar to part (b), but with each die having a $\frac{18}{20}$ probability of not coming up 1 or 20. So the probability of that case is $(\frac{18}{20})^{20}$.

For Case ii., we can further subdivide this (into mutually exclusive and exhaustive subcases) based on where in the sequence we see our first 1:

- If we get a 1 on the first die, then it doesn't matter what the remaining dice are (they could even be more 1s), as long as they are not 20. The probability of this is $\frac{1}{20} \cdot (\frac{19}{20})^{19}$.
- We could get something other than 1 or 20 as the first die, then get 1 as the second die. Then it doesn't matter what the remaining dice are, as long as they are not 20. The probability of this is $\frac{18}{20} \cdot \frac{1}{20} \cdot (\frac{19}{20})^{18}$.

And so on. An expression for the sum of these probabilities is $\sum_{i=0}^{19} (\frac{18}{20})^i (\frac{1}{20}) (\frac{19}{20})^{19-i}$.

Case iii. can be handled the same way as Case ii. Therefore the probability of Case iv., which is what we want, is 1 minus the sum of the other three Cases, i.e.,

$$\boxed{1 - (\frac{18}{20})^{20} - 2\sum_{i=0}^{19} (\frac{18}{20})^i (\frac{1}{20}) (\frac{19}{20})^{19-i}}.$$

As is often the case in combinatorics, there may be a nicer way to do this! Let me know if you find one.

If you got a different expression and want to check it on Wolfram Alpha: that answer is

Result

$$\frac{21212944650652080893863087}{52428800000000000000000000} \approx 0.404605$$

(e) To explain the first expression:

- We will consider the rolls as an ordered string.
- There are 6 possible outcomes for each die, so there are 6^5 possible outcomes overall. This is the sample space.
- We have 6 choices for the number that will be the group of 3, and 5 choices for the number that will be the group of 2.

- We have $\binom{5}{2}$ ways of assigning the two rolls that will be our "group of 2" number.

So what goes wrong when we try to apply the same reasoning to the four-die case?

- We will consider the rolls as an ordered string. - still OK
- There are 6 possible outcomes for each die, so there are 6^4 possible outcomes overall. This is the sample space. - still OK
- We have 6 choices for the number that will be the group of 2, and 5 choices for the number that will be the group of 2. – oh no! Now the groups are of equal size, so saying "the first group of 2 will be 3s and the second group of 2 will be 4s" is the same as saying "the first group of 2 will be 4s and the second group of 2 will be 3s". So we are double-counting – an outcome like 4334 will be counted once as "two fours and two threes", and again as "two threes and two fours"! What we really want here is 'one group of 2 will be 4s and the other group of 2 will be 3s". We can fix the double-counting by dividing by 2.

Another equivalent way to frame the answer is as $\frac{\binom{6}{2}\binom{4}{2}}{6^4}$. We just need to choose two of the six possible numbers to be our repeated values, but there is no notion that one is "first" and one is "second".