EECS 70 Discrete Mathematics and Probability Theory
Fall 2014 Anant Sahai

Homework 14

This homework is due December 9, 2014, at 12:00 noon.

1. Section Rollcall!

In your self-grading for this question, give yourself a 10, and write down what you wrote for parts (a) and (b) below as a comment. Please put the answers in your written homework as well.

(a) What discussion did you attend on Monday last week? If you did not attend section on that day, please tell us why.

(b) What discussion did you attend on Wednesday last week? If you did not attend section on that day, please tell us why.

2. Practice Makes Perfect

For this question, do 5 of the online practice problems. For your answer, write down which problems you did (the problem set title and the number of the question). Use a screen capture to show us that you finished them.

3. Random Variables and Distributions Lab

In this week's lab, we will explore common discrete random variables and their corresponding distributions.

For each part, students who want to can choose to completely rewrite the question. Basically, you can come up with your own formulation of how to do a series of experiments that result in the same discoveries. Then, write up the results nicely using plots as appropriate to show what you observed. You can also rewrite the entire lab to take a different path through as long as they convey the key insights aimed at in each part.

Please download the IPython starter code from Piazza or the course webpage, and answer the following questions.

(a) Plot the PMFs of binomial random variables with $N = 20$, and with success probabilities $p = 0.1, 0.3, 0.5, 0.7$, and 0.9, respectively. You should have 5 different plots in one figure.

What do you observe as the success probability increases?

(b) In probability theory and statistics, the cumulative mass function (CMF) describes the probability that a real-valued discrete random variable X with a given probability distribution will be found to have a value less than or equal to x. Mathematically, we define the CMF $F_X(x)$ as

$$F_X(x) = P(X \leq x),$$

where the right-hand side represents the probability that the random variable X takes on a value less than or equal to x.

Plot the CMFs of the five binomial random variables from part (a). These should all be increasing curves. At what value does each CMF plot converge to and stay at 1 (i.e. at what value can you be almost 100% certain that the number of heads (or successful trials) is less than such value)?

(c) Plot the PMF a binomial distribution with parameters ($N = 100, p = 0.2$) in a bar chart. Then, overlay your plot with the probability density function (PDF) of a normal distribution with parameters ($\mu = 20, \sigma^2 = 16$). What do you observe?

In a different figure, plot the CMF of the same binomial distribution and overlay it with the cumulative density function (CDF) of the aforementioned normal distribution. Again, what do you observe?

Finally, derive an approximation between the two distributions using a concept you learned in this week's lecture.

There are other optional parts of this Virtual Lab in the file `vl14.pdf`, which you can find on Piazza or the course website.

Reminder: When you finish, don't forget to convert the notebook to pdf and merge it with your written homework. Please also zip the `ipynb` file and submit it as `hw14.zip`.

4. To Be "Normal"

Suppose a standard 6-sided die (with faces 1 through 6) is rolled n times, and let A be the average of the results.

(a) How does the Central Limit Theorem help us approximate the distribution of A?

(b) Let A' be a random variable drawn from the Gaussian distribution best approximating the distribution of A. If $n = 100$, what are the bounds of an interval $[a,b]$ centered at 3.5 such that $A' \in [a,b]$ with probability exactly 90%?. (See `https://statistics.laerd.com/statistical-guides/normal-distribution-calculations.php` for normal distribution calculation. You might want to use Table 1.)

(c) Approximate the probability that $3 \leq A' \leq 4$, if $n = 30$.

(d) What is the minimum n for which, with probability at least 99%, we have $3 \leq A' \leq 4$?

5. Hypothesis testing

We would like to test the hypothesis claiming that a coin is fair, i.e. $P(H) = P(T) = 0.5$. To do this, we flip the coin $n = 100$ times. Let Y be the number of heads in $n = 100$ flips of the coin. We decide to reject the hypothesis if we observe that the number of heads is less than $50 - c$ or larger than $50 + c$. However, we would like to avoid rejecting the hypothesis if it is true; we want to keep the probability of doing so less than 0.05. Please determine c. (*Hints: use the central limit theorem to estimate the probability of rejecting the hypothesis given it is actually true.*)

You might need to use Table 1.

6. Simplified Self-Grading (carried over from HW13; only parts (d)-(h) need to be done)

There are about $n = 500$ self-graded question parts in this iteration of EECS 70. For this simplified version of self-grading, we use a scale from 0 to 4 instead of the $0, 2, 5, 8, 10$ scale currently being used. On each of them, a student assigns a grade S_i. For each homework, readers randomly grade a subset of the problems. Assume that $n/5$ of the question parts are graded by the readers (chosen uniformly over all the problem parts) and the readers assign grades R_i. Assume that R_i may deviate from an honest self-grade S_i according to the conditional probabilities given in Table 2.

We do the following check: we add up all of the $S_i - R_i$ for a particular student (for the subset of problems graded by readers only). If the result is too high, we suspect that a student might be inflating their grades.

Probability Content from -oo to Z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Table 1: Table of the Normal Distribution.
Source: `http://cosstatistics.pbworks.com/w/page/27425647/Lesson`

R_i \ S_i	0	1	2	3	4
0	3/4	1/4	0	0	0
1	1/4	1/2	1/4	0	0
2	0	1/4	1/2	1/4	0
3	0	0	1/4	1/2	1/4
4	0	0	0	1/4	3/4

Table 2: $\mathbf{P}(R_i|S_i)$.

(a) Suppose that a student is honest. Let $p_0 = \mathbf{P}(S_i = 0)$ and $p_4 = \mathbf{P}(S_i = 4)$. Let $X_i = S_i - R_i$. Express the distribution of X_i as a function of p_0 and p_4.

(b) Give the best upper-bounds you can on both $\mathbf{E}[X_i]$ and $\mathrm{Var}(X_i)$. Your bounds shall not depend on p_0 or p_4.

(c) Using Chebyshev's inequality and the above parts, compute the smallest threshold T that we should

choose so that $\sum_i X_i \leq T$ for 95% of honest students?

(d) Repeat the above using the Central Limit Theorem to get an approximate answer for T. (You might want to refer to Table 1 for the cumulative normal distribution table.)

(e) For simplicity, we are going to focus our attention on a hypothetical student who never truly deserves full points and never truly deserves a zero on any question part, i.e., they never give themselves a zero or full points on a question. Recompute better upper bounds on both $\mathbf{E}[X_i]$ and $\text{Var}(X_i)$ that are valid for this student. Recalculate the relevant theshold T using the Central Limit Theorem.

(f) Assume this student is inflating their true self-grades S_i by adding 1 point to a question part with probability $1/2$. What is their risk of being caught (i.e., above the threshold T)? (Here, explain how you are modeling things to be true to the spirit of this problem.)

(g) If this student is willing to accept a 50% chance of being caught cheating, by how much can they systematically inflate their grade(i.e. inflates his/her grade to every question by some constant number of points)? Assume that they can inflate by no more than 3 points per question part. (Because inflating a 0 to a 4 would get them slammed the first time they did it.) We will assume that $5, 6$ and 7 are allowed as self-reported grades to keep things simple.

(h) Is it worth trying to cheat on self-grading, even for a grade-maximizing sociopath[1] student with no internal sense of morality or "decent respect to the opinions of mankind." ?

7. Tolerating Errors

Assume Alice is trying to send m packets across a noisy channel to her friend Bob. The channel independently has probability p of generating an error on each packet. To account for errors, Alice sends $n > m$ packets. If Alice wants to ensure that Bob can correctly decode her entire message of m packets with probability at least r, how big can m be?

(a) Modeling each error as a coin toss with probability p, what is the probability that Bob cannot correctly decode Alice's message?

(b) Assume $n = 100$, $r = 0.9$, and $p = 0.1$. What is the safe bound for m using the Chebyshev bound?

(c) Using the same information as part (b), approximate the safe m using the Central Limit Theorem.

(d) Using the same information as part (b), what is the safe bound for m using the relevant Chernoff bound for Binomial RV's?

8. Wrecking Ball

A new startup Milton/Alighieri Games has decided to create a family of games to cater to a previously underserved segment of the video game market. In their quest for a crossover hit, they score a marketing coup by getting the exclusive video game rights to use Miley Cyrus's megahit song, "Lucifer's Lament" which RCA Records had been promoting solely under its alternate title to avoid drawing controversy[2]. This family of game titles are all set in the mythical "War/Rebellion in Heaven" and include real time strategy games (like Starcraft), a Multiplayer Online Battle Arena game (like League of Legends), and an online Collectible Card Game (like Magic: The Gathering or Hearthstone).

[1] This sociopathic model of a selfish maximizer is referred to as a "rational agent" in the formal language of economics. Showing that cheating is not substantially attractive in the context of a mechanism even for a sociopath is one way to show that the mechanism is probably safe against normal humans too — since actual human beings are caring, loving, altruistic, and have senses of integrity and honor.

[2] Miley had grown increasingly annoyed since many people weren't appreciating the emotional nuances of her performance without seeing the mythic allusions resonating with it.

You're in charge of one of the main characters, and are proposing a particular attack that you've entitled "Wrecking Ball" to tie in with the song. In this attack, at every turn, one of two things happens: With probability $\frac{2}{3}$, the attack succeeds and the character's health levels double (raising your max if necessary), with that many health points being drained from the opponent. With probability $\frac{1}{3}$, the attack backfires and the character's health gets reduced by a factor of 8, with the lost health being transfered to the opponent. Different turns are independent in whether the attack succeeds or backfires.

This problem is about understanding what happens if this attack is used repeatedly.

Let X_0 be the initial real-valued health of the character. So the health at the end of turn n is $X_n = X_0 \prod_{i=1}^{n} A_i$ where A_i is the random factor that results from the attack in turn i.

(a) Calculate the expected value of A_i.

(b) Calculate the expected value of X_n. You can assume that $X_0 = x_0$ — a given constant.

(c) Seeing the previous calculations, the reasonably pious management gets very concerned about your proposed attack and worry that perhaps it is tilting the game in favor of this character.
Explain to them whether they are right or wrong to be worried by explaining to them what the typical range for X_n should be (with probability at least 90%) when n is large.
(Hint: figure out how to invoke the laws of large numbers.)

9. A Chernoff Bound

In this problem, you will show that the probability that the average of specific 3-valued independent random variables is "far away" from its expectation decays exponentially in the number of random variables in the sum. You have already seen how to do this for independent Bernoulli random variables via the Chernoff bound in the notes; now, we will explore how to do this for the following independent 3-valued random variables. Let X_i be a sequence of independent identically distributed random variables such that:

$$X_1 = \begin{cases} 2 & \text{with prob. } 1/3 \\ 1 & \text{with prob. } 1/3 \\ 0 & \text{with prob. } 1/3 \end{cases}$$

Let $X = \sum_{i=1}^{n} X_i$.

a) Argue why $\mathbf{P}(X \geq na) = \mathbf{P}(e^{sX} \geq e^{nsa})$ for all values $s \geq 0$. Our goal is to come up with an upper bound for this quantity that decays exponentially with n. Argue why we should only concern ourselves with $1 < a < 2$ (Is $a \leq 1$ interesting for this bound?).

b) Argue why $\mathbf{P}(X \leq na) = \mathbf{P}(e^{sX} \geq e^{nsa})$ for all values $s \leq 0$. Argue why we should only concern ourselves with $0 \leq a < 1$.

c) Since it was the right hand side of both the above equalities, let's focus on bounding $\mathbf{P}(e^{sX} \geq e^{nsa})$. Show that $\mathbf{P}(e^{sX} \geq e^{nsa}) \leq e^{-n(sa - \ln M(s))}$, where $M(s) = \mathbf{E}[e^{sX_i}]$.

d) Compute $M(s) = \mathbf{E}[e^{sX_i}]$.

e) Now we have the tools to continue part a) and start to bound $\mathbf{P}(X \geq na)$. Use the parts above to conclude that $\mathbf{P}(X \geq na) = \mathbf{P}(e^{sX} \geq e^{nsa}) \leq e^{-n \max_{s \geq 0}(sa - \ln M(s))}$.

f) Plug $s = 0$ into $sa - \ln M(s)$. Given this value, what can you conclude about the *maximum* value of $sa - \ln M(s)$ for $s \geq 0$?

g) Give the value of s that maximizes $sa - \ln M(s)$ for $s \geq 0$. Show that this is a positive value for s given that $1 < a < 2$. What does the fact that this is a *positive* value for s tell you about the value of $sa - \ln M(s)$ when maximized over $s \geq 0$? Potentially Useful Hint: If you want to show that $(x+y)/z > 1$, you can start by determining whether x, y and z are nonnegative or negative. Once you know this, you can manipulate the inequality $(x+y)/z > 1$ to get an equivalent inequality that you can verify more easily. Also, if α and β are positive, then $\alpha > \beta \iff \alpha^2 > \beta^2$.

h) Give an upper bound that decays exponentially with increasing n for $\mathbf{P}(X \geq na)$, using your previous parts to justify it.

i) **(Optional)** Now complete the exponential upper bound for b). In part b), $s \leq 0$ and we want to bound $\mathbf{P}(X \leq na)$. With this in mind, repeat similar arguments to those in the previous parts to come up with a bound that decays exponentially with n for $\mathbf{P}(X \leq na)$, where $a < 1$.

j) **(Virtual Lab, Optional)** For $a = 1.5$, do an appropriate simulation using a computer and plot the actual probability of having this sort of large deviation happen as compared to what your bounds above say. Use the appropriate kind of axes to judge the quality of the bound.

10. (Optional) 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.

(a) Compute the mean and variance of X. Show that $\mu = E[X] = n/2$ and $\sigma^2 = \text{Var}[X] = n/4$.

(b) Prove that $\Pr[X = k] = \binom{n}{k}/2^n$.

(c) Show by using Stirling's formula that $\Pr[X = k] \simeq \frac{1}{\sqrt{2\pi}}(\frac{n}{2k})^k(\frac{n}{2(n-k)})^{n-k}\sqrt{\frac{n}{k(n-k)}}$.

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 $\Pr[Y = t]/d = \Pr[X = k]/d$.

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 $\Pr[Y = t]/d$. Show that the end result is equivalent to

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

(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 $\exp(\ln(1+x)(1+x))$ and then replace $\ln(1+x)(1+x)$ by its Taylor series.
The Taylor series up to the x^2 term is $\ln(1+x)(1+x) \simeq x + x^2/2 + \ldots$ (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:

$$\frac{1}{\sqrt{2\pi}}e^{-t^2/2}$$

(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.)

11. Write your own problem

Write your own problem related to this week's material and solve it. You may still work in groups to brainstorm problems, but each student should submit a unique problem. What is the problem? How to formulate it? How to solve it? What is the solution?