CS 70 Discrete Mathematics for CS
Spring 2008 David Wagner

HW 13

Due Thursday, March 8th

Important: Show your work on all problems on this homework.

1. (6 pts.) The myth of fingerprints, cont.

A crime has been committed. The police discover that the criminal has left DNA behind, and they compare the DNA fingerprint against a police database containing DNA fingerprints for 20 million people. Assume that the probability that two DNA fingerprints (falsely) match by chance is 1 in 10 million. Assume that, if the crime was committed by someone whose DNA fingerprint is on file in the police database, then it's certain that this will turn up as a match when the police compare the crime-scene evidence to their database; the only question is whether there will be any false matches.

Let D denote the event that the criminal's DNA is in the database; $\neg D$ denotes the event that the criminal's DNA is not in the database. Assume that it is well-documented that half of all such crimes are committed by criminals in the database, i.e., assume that $\Pr[D] = \Pr[\neg D] = 1/2$. Let the random variable X denote the number of matches that are found when the police run the crime-scene sample against the DNA database. In the following, compute all probabilities to at least two digits of precision.

(a) Calculate $\Pr[X = 1|D]$.

(b) Calculate $\Pr[X = 1|\neg D]$.

(c) Calculate $\Pr[\neg D|X = 1]$.

(d) Suppose that the police find exactly one match, and promptly prosecute the corresponding individual. Suppose you are appointed a member of the jury, and the DNA match is the only evidence that the police present. Do you think the defendant should be convicted? Why or why not?

2. (6 pts.) Chebyshev inequality

A friend tells you about a course called "Laziness in Modern Society" that requires almost no work. You hope to take this course so that you can devote all of your time to CS70. At the first lecture, the professor announces that grades will depend only a midterm and a final. The midterm will consist of three questions, each worth 10 points, and the final will consist of four questions, also each worth 10 points. He will give an A to any student who gets at least 60 of the possible 70 points.

However, speaking with the professor in office hours you hear some very disturbing news. He tells you that to save time he will be grading as follows. For each students midterm, he'll choose a real number randomly from a distribution with mean $\mu = 5$ and variance $\sigma^2 = 1$. He'll mark each of the three questions with that score. To grade the final, he'll again choose a random number from the same distribution, independent of the first number, and he'll mark all four questions with that score.

If you take the class, what will the mean and variance of your total class score be? Can you conclude that you have less than a 5% chance of getting an A? Why?

3. (6 pts.) Normal distribution

Suppose the overall scores of the students in a Discrete Math class are approximately normally distributed with a mean of 83 and a standard deviation of 6. Compute:

(a) the lowest passing score, if the bottom 5% of these students fail;

(b) the highest B, if the top 10% of the students are given A's.

Note: You may assume that if X is normal with mean 0 and variance 1, then $\Pr[X \leq 1.3] \approx 0.9$ and $\Pr[X \leq 1.65] \approx 0.95$.

4. (9 pts.) z-scores

Let Z be a random variable that has a normal distribution with mean 0 and variance 1. Given a real number z, there are tables that allow us to compute $\Pr[Z \leq z]$ as a function of z. The value z is sometimes called a z-score. The tables allow us to compute one (or both) of the following quantities:

- The "left tail": The left tail represents the values of Z that are less than or equal to z, and $\Pr[Z \leq z]$ is the area under the normal curve and where $x \leq z$. For instance, $\Pr[Z \leq -1] \approx 0.1587$, $\Pr[Z \leq 0] = 0.5$, and $\Pr[Z \leq 1] \approx 0.8413$.
- The "right tail": The right tail represents the values greater than or equal to z. For instance, $\Pr[Z \geq 1] \approx 0.1587$, and $\Pr[Z \geq 2] \approx 0.0228$.

You can find resources for calculating these values—e.g., normal tables, normal calculators—at `http://www.cs.berkeley.edu/~daw/teaching/cs70-s08/tables.html`. These typically allow you to go back and forth between a z-score z and the probability $\Pr[Z \leq z]$ (i.e., the area under the "left tail"), or between z and $\Pr[Z \geq z]$ (i.e., the area under the "right tail").

(a) Let the r.v. Z be normally distributed with mean 0 and variance 1. Use a table or calculator mentioned above to find the approximate value of $\Pr[Z \leq 1.5]$.

z-scores have many applications. For instance, if the random variable X is normally distributed with mean μ and variance σ^2, then the random variable X can be *normalized* to get a random variable X_{norm} defined by $X_{\text{norm}} = (X - \mu)/\sigma$. A useful fact is that, with these assumptions, X_{norm} will be normally distributed with mean 0 and variance 1. Note that the value of X_{norm} can be viewed as a z-score.

(b) Suppose X is normally distributed with mean 100 and standard deviation 10. Calculate $\Pr[X \geq 125]$. You may wish to use the resources listed above.

Here is another application of z-scores. Let B be the number of Heads after flipping n coins, with Heads probability p, i.e., $B \sim \text{Binomial}(n, p)$. We have shown that $\mathbf{E}[B] = np$ and $\text{Var}(B) = np(1-p)$. It turns out that, for large n, the binomial distribution B approximates the normal distribution with the same mean and variance. Let's normalize B, to get a random variable B_{norm} defined as follows:

$$B_{\text{norm}} = \frac{B - np}{\sqrt{np(1-p)}}.$$

Given the assumption that B is approximately normally distributed with mean np and variance $np(1-p)$, then B_{norm} is approximately normally distributed with mean 0 and variance 1. Thus, the value of B_{norm} can be viewed as a z-score.

(c) Find a value k for which, when you flip a fair coin 10,000 times, the probability of k or more heads is approximately 0.20.

5. (8 pts.) Statistical significance

A college soccer team won its conference championship 11 times in the first 20 years of existence. Then, for the next 20 years it won only 3 times. Prof. Argyle argues that the soccer team has gotten worse; but Prof. Plaid insists that this difference is just due to bad luck. Let's figure out whether this difference is statistically significant or whether it can be plausibly attributed to random chance.

To model this, let's suppose the probability of winning the championship was p in each of the first 20 years, and q in the each of the second 20 years, and each year's results are independent of all other years. Define the r.v. $X =$ the total number of wins in the first 20 years, $Y =$ the total number of wins in the next 20 years, and $Z = X - Y$. Under these assumptions, we'd have $X \sim \text{Binomial}(20, p)$ and $Y \sim \text{Binomial}(20, q)$. Using the Normal approximation, we can conclude that X and Y have approximately a Normal distribution, and consequently $Z = X - Y$ must also have an approximately Normal distribution. We don't know what p and q are, but we're going to try to develop a statistical test to check whether the observation $Z = 11 - 3 = 8$ seems to be plausibly consistent with the hypothesis that $p = q$.

(a) Under the model of the second paragraph, and assuming that $p = q$, compute $\mathbf{E}[Z]$.

(b) Under the model of the second paragraph, and assuming that $p = q$, what is the largest that $\text{Var}(Z)$ could be?

(c) Under the model of the second paragraph, and assuming that $p = q$, compute an upper bound for $\Pr[Z \geq 8]$ and evaluate it to at least 2 digits of precision.

Hint: Use the Normal approximation for Z, and use z-scores.

(d) Now let's try to decide whether it seems plausible that Prof. Plaid is right, given the information that $Z \geq 8$. In the following, we'll use a 95% confidence level. We'll compute the probability that $Z \geq 8$, assuming that Prof. Plaid is right (i.e., given that $p = q$). If this probability is less than 0.05, we'll decide that the observed value $Z = 8$ seems too implausible to have happened by chance, we'll call the difference statistically significant (at 95% confidence level), and we'll reject Prof. Plaid's hypothesis (with 95% confidence). On the other hand, if the probability is above this threshold, we'll say that the data is inconclusive, the difference is statistically insignificant, and we cannot rule out Prof. Plaid's claim—he might be right.

So, in the given example where we observe $Z = 8$, what's the bottom line? Is this difference statistically significant, or not? Can we reject Prof. Plaid's hypothesis, at 95% confidence level?

6. (15 pts.) Those 3407 Votes

In the aftermath of the 2000 US Presidential Election, many people have claimed that unusually large number of votes cast for Pat Buchanan in Palm Beach County are statistically highly significant, and thus of dubious validity. In this problem, we will examine this claim from a statistical viewpoint.

The total percentage votes cast for each presidential candidate in the entire state of Florida were as follows:

Gore	Bush	Buchanan	Nader	Browne	Others
48.8%	48.9%	0.3%	1.6%	0.3%	0.1%

In Palm Beach County, the actual votes cast (before the recounts began) were as follows:

Gore	Bush	Buchanan	Nader	Browne	Others	Total
268945	152846	3407	5564	743	781	432286

To model this situation probabilistically, we need to make some assumptions. Let's model the vote cast by each voter in Palm Beach County as a random variable X_i, where X_i takes on each of the six possible

values (five candidates or "Others") with probabilities corresponding to the Florida percentages. (Thus, e.g., $\Pr[X_i = \text{Gore}] = 0.488$.) There are a total of $n = 432286$ voters, and their votes are assumed to be mutually independent. Let the r.v. B denote the total votes cast for Buchanan in Palm Beach County (i.e., the number of voters i for which $X_i = \text{Buchanan}$).

(a) Compute the expectation $\mathbf{E}[B]$ and the variance $\text{Var}(B)$.

(b) Use Chebyshev's inequality to compute an *upper bound* b on the probability that Buchanan receives at least 3407 votes, i.e., find a number b such that

$$\Pr[B \geq 3407] \leq b.$$

Based on this result, do you think Buchanan's vote is significant?

(c) Suppose that your bound b in part (b) is exactly accurate, i.e., assume that $\Pr[X \geq 3407]$ is exactly equal to b. [*In fact the true value of this probability is much smaller, as you shall see in part (e).*] Suppose also that all 67 counties in Florida have the same number of voters as Palm Beach County, and that all behave independently according to the same statistical model as Palm Beach County. What is the probability that in *at least one* of the counties, Buchanan receives at least 3407 votes? How would this affect your judgement as to whether the Palm Beach tally is significant?

(d) Our model assumes that all voters behave like the fabled "swing voters," in the sense that they are undecided when they go to the polls and end up making a random decision. A more realistic model would assume that only a fraction (say, about 20%) of voters are in this category, the others having already decided. Suppose then that 80% of the voters in Palm Beach County vote deterministically according to the state-wide proportions for Florida, while the remaining 20% behave randomly as described earlier. Does your bound b in part (b) increase, decrease or remain the same under this model? Justify your answer.

(e) Now let's use the Normal approximation for B to compute a better estimate to $\Pr[B \geq 3407]$, under the same "swing voter" model used in part (b). How many standard deviations away from the mean is this observed value for B? Explain why it's difficult to use the normal tables listed in problem 4 to compute the probability that a Normally distributed r.v. is this far away from the mean.

7. (0+5 pts.) Optional bonus problem

Disturbed by the solutions to the last homework set, a prison warden decides to give all 101 inmates in his county jail a chance at leniency: he offers them a chance to play a game. The warden has an inexhaustible supply of blue hats and red hats, and he will secretly flip a fair coin for each inmate to select that inmate's hat color. Each inmate can see the color of all other inmates hats, but cannot see his or her own hat color. The inmates are allowed to strategize in advance, but once the hats go on, no communication of any form is allowed. After everyone receives their hats, each inmate receives a slip of paper and must write down a guess at whether the total number of blue hats (including his or her own hat) is odd or even. If at least 51 of the inmates guess correctly, everyone wins and all the inmates are set free. Otherwise, no one wins.

Obviously, the inmates can ensure a 50% chance of freedom just by having each inmate guess randomly. But can you do better than 50%? What's the best strategy you can come up with? What is the probability they go free, under your strategy?