Assignment 3 Solutions

CSCI2244-Randomness and Computation
Due Thursday, February 8, at 11:30PM

1 Poker

If I asked you to compute the probabilities of various poker hands, it would take you just a few seconds find the Wikipedia page 'Poker odds' with all the answers, complete with the number of relevant outcomes for each hand expressed in terms of binomial coefficients. Please do take a look at it! An excellent exercise is to try to find the correct reasoning behind the results they give, and obtain the same answers.

But I needed to come up with problems whose answers you can't look up on the Web. So I invented some new poker hands, and here I ask you to determine their probabilities. This is the version of poker where we draw 5 cards from a shuffled deck. As in our examples in class, we model this problem with the sample space

$$\{X \subseteq \{1, \ldots, 52\} : |X| = 5\},$$

and assume each 5-element subset has the same probability.

Explain your reasoning carefully, and express your answers both as a formula using binomial coefficients and powers, and as numerical values. It's easy to be led astray here, and a very good way to check your answer is to write a simulation. You are not required to do this for the homework, but it's not a bad idea if you want to see if you were right.

(a) *A blush* All 5 cards are red.

Solution. There are 26 red cards, and thus the number of blushes is $\binom{26}{5.}$ So the probability of a blush is

$$\frac{\binom{26}{5}}{\binom{52}{5}} \approx 0.0253.$$

(b) *A flash.* All four suits occur in the hand. (This means that one suit will occur twice, and the others once, which is a hint for solving the problem. The problem gets more complicated if you attempt to filter out flashes that belong to some other category of poker hand, like pairs or straights. So just assume that any hand containing all four suits is a flash.)

Solution. This is the hardest one. We count the number of flashes as a sequence of choices. We first choose the suit which will occur twice. There are four ways to do this. We then choose which two cards of this suit to include. There are $\binom{13}{2}$ ways to do this. Finally, for each of the three remaining suits, there are 13 ways to choose the card in that suit. The result is

$$4 \cdot \binom{13}{2} \cdot 13^3 = 685464,$$

so the probability is

$$\frac{4 \cdot \binom{13}{2} \cdot 13^3}{\binom{52}{5}} \approx 0.264.$$

(c) *A royal scandal.* The hand contains two Kings, two Queens, and a Jack.

Solution. We count the number of royal scandals by treating each hand as the result of a sequence of 3 choices: $\binom{4}{2}$ ways to choose the Kings, $\binom{4}{2}$ ways to choose the Queens, and 4 ways to choose the Jacks. So the probability is

$$\frac{\binom{4}{2}^2 \cdot 4}{\binom{52}{5}} \approx 5.54 \times 10^{-5}.$$

2 Elections

There are two candidates in an election. Candidate A has received 55% of the votes, candidate B 45%. There is a very large number of voters (several million, let's say). We randomly sample 100 voters. What is the probability that in this sample, candidate B receives more votes? (In other words, that the poll does not correctly predict the outcome of the election.) Express this answer as a formula using the binomial coefficients, and then compute the numerical value.

HINT: Strictly speaking, this is sampling without replacement, since we should not poll the same voter twice! But the voter pool is so large and the sample so

small in comparison, that you can treat it as a problem of sampling with replacement, which makes the calculation somewhat easier. You can then think of the problem as one of flipping 100 biased coins in succession. We saw how to express the probability of getting exactly k heads in terms of binomial coefficients, so here you will have a sum of about 50 such probabilities. To obtain a numerical value you will need to write a little code.

Solution. The event in question is

$$\bigcup_{k=51}^{100} E_k,$$

where E_k is the event 'candidate B gets k votes'. These events are pairwise disjoint, so the probability is

$$\sum_{k=51}^{100} P(E_k).$$

$P(E_k)$ is the probability of getting k heads on 100 tosses of a coin with heads probability 0.45; that is,

$$\binom{100}{k} 0.45^k 0.55^{100-k}.$$

So the answer is

$$\sum_{k=51}^{100} 0.45^k 0.55^{100-k}.$$

Later we will see how to estimate this with so-called normal distribution. Python will compute it with a single line:

```
>>> sum([sp.binom(100,k)*0.45**k*0.55**(100-k)for k in range(51,101)])
0.13457621318804303
```

Put another way, 87% of the time, the sample of 100 people correctly predicts the election if one candidate got at least 55% of the vote.

3 The birthday party never ends!

I can't seem to leave the birthday stuff alone, and I had to restrain myself from making *every* problem about birthdays.

3.1 *Somebody* has a birthday today.

When we did the experiment in class on January 25 with a group of 30 people, we found that there was a student whose birthday was that very day. Now, because of our calculations, I knew going in that I would probably find two students with the same birthday. But what is the probability that I would find a student with a birthday on January 25? Give the answer both for 30 students, and in general for k students. How many students would have to be present to make this probability at least $\frac{1}{2}$? You can solve this problem exactly, or use the exponential approximation.

Solution. We use the same probability model as in our original solution of the birthday problem, but we look at the complementary event 'no one has a January 25 birthday'. The same reasoning shows that the probability we want is

$$1 - \frac{364^{30}}{365^{30}} = 1 - \left(\frac{364}{365}\right)^{30} \approx 0.079.$$

How many people do we need to have this probability equal to $\frac{1}{2}$? We have to solve

$$\left(\frac{364}{365}\right)^{k} = \frac{1}{2}.$$

We take logarithms (base doesn't matter) of both sides and do a bit of rearranging to get

$$k = \frac{\log 2}{\log 365 - \log 364} \approx 252.65,$$

so we need at least 253 people present.

Alternatively, we can compute this approximately, using the exponential approximation

$$\frac{364}{365} \approx e^{\frac{-1}{365}},$$

so we end up solving

$$e^{\frac{-k}{365}} = \frac{1}{2}.$$

Taking natural logs of both sides gives

$$k = 365 \cdot \ln 2 = 252.99.$$

That gives the same result in terms of a whole number of people as the exact solution, but is somewhat easier to compute.

3.2 Births and deaths

Suppose you have a database of biographies of prominent people from the past. Each biography contains a date of birth and a date of death. If there are 1000 records in the database, what is the probability that two of them share both a date of birth and a date of death (we are ignoring the year of birth and the year of death, and just looking at the month and the day)? You should use the exponential estimate for the generalized birthday problem.

It is exactly the same computation as the one we made in class for a coincidental birthday in a group of 30 people, but we just replace 30 by 1000, and 365 by 365^2, which is the number of pairs of dates. We wind up having to evaluate

$$1 - \exp(-(1000 \cdot 999)/(2 \cdot 365^2)),$$

which is again one line of Python:

```
>>> 1-math.exp(-(999*1000)/(2*365*365))
0.9816843611112658
```

It's just about a sure thing.

3.3 Real birthdays

As we discussed, our calculation of the probability of a shared birthday in a group of people was based on some idealized assumptions: First, we treated the problem as one of sampling with replacement, but that is not such a big deal (see the elections problem above). Second, there are actually 366 possible birthdays, but one of them (February 29) occurs much less frequently than the others. Third, we assumed that that the distribution of the 365 birthdays is uniform—but are births really uniformly distributed throughout the year?

To help you answer the question, I have posted data on US births in the years 2000-2003. (This was extracted from files posted on GitHub by FiveThirtyEight.com, who in turn got it from records of the Social Security Administration.) I modified the file format ever so slightly, and retained just these four years—the original data was for the 15 years from 2000 to 2014. Thus the data given here contains one leap year and three non-leap years, so February 29 is represented about the right amount.

This is a '.csv' file, which means that it is actually an ordinary text file, with each line consisting of several fields (5 in this case) separated by commas. If you

double-click on it, it will probably open with Excel, and, indeed, you could do part (a) below just using Excel stuff. But you will need to open and read the file with Python to do part (b), so I've included in an appendix instructions about how to do this.

(a) Make two scatter plots or stem plots of the data, showing the proportion of the total number of births for each day of the year, both for the year 2000 alone, and for all four years combined. Note that there will be 366 x-values, and you will need to be careful because February 29 only occurs during the first year. (As a reality check, the plot for 2000-2003 should show February 29 as an outlier, with many fewer births, but for 2000 it will appear as a 'normal' day.)

In both plots you can see the seasonal nonuniformity very clearly. In the plot for a single year, there is another source of non-uniformity, which I find absolutely astonishing and somewhat perplexing: It looks as if there are two or even three entirely separate data series, with roughly the same seasonal variation, but one significantly lower than the other.

You won't be graded on this, but you are invited to speculate on the reasons for these non-uniformities.

(b) You are to estimate, for $k = 1$ to 65, the probability of a coincidental birthday in group of k people, by performing a simulation based on the probability distribution you computed in (a) for all four years. You should then superimpose this on the plot that computes these probabilities under the assumption of a uniform distribution of 365 days. (The code for this is on the website.) Do you see much difference between the two plots? How adequately does the uniform probability distribution model the real-life version of this problem?

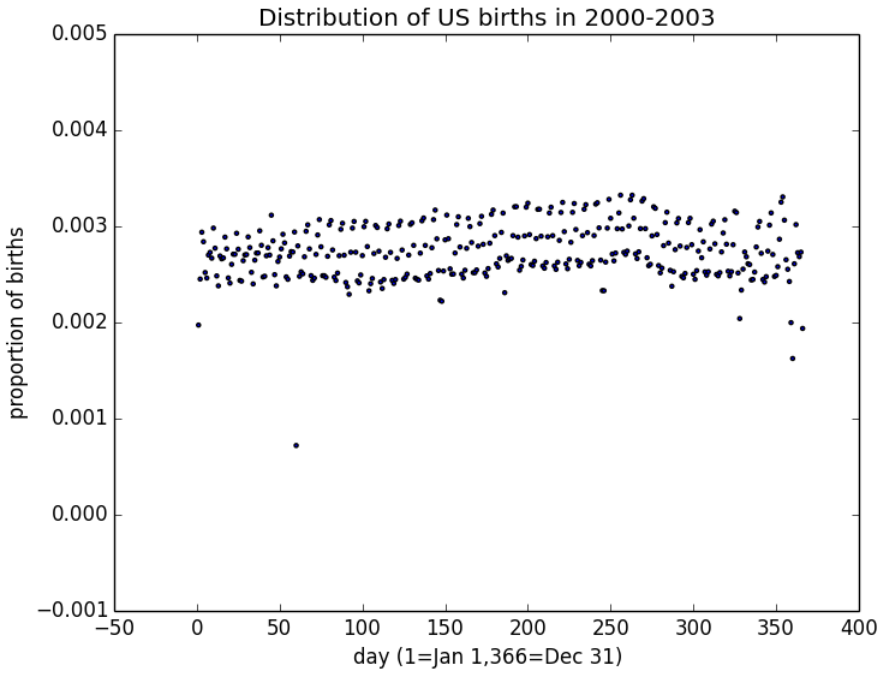
Solution. The accompanying code contains the routines for extracting the data from the file and producing the scatter plots, both for the year 2000 alone and the years 2000-2003 combined, giving the true distribution of the 365 birthdays for this time period, as well as the code for plotting the probability of coincidental birthdays. As it turns out, I used the more straightforward simulation method, sampling for each number of people in the room–this did not take all that long. (In some respects a stem plot would have been better, because you could more easily see which date corresponds to which dot.) The plots are shown below.

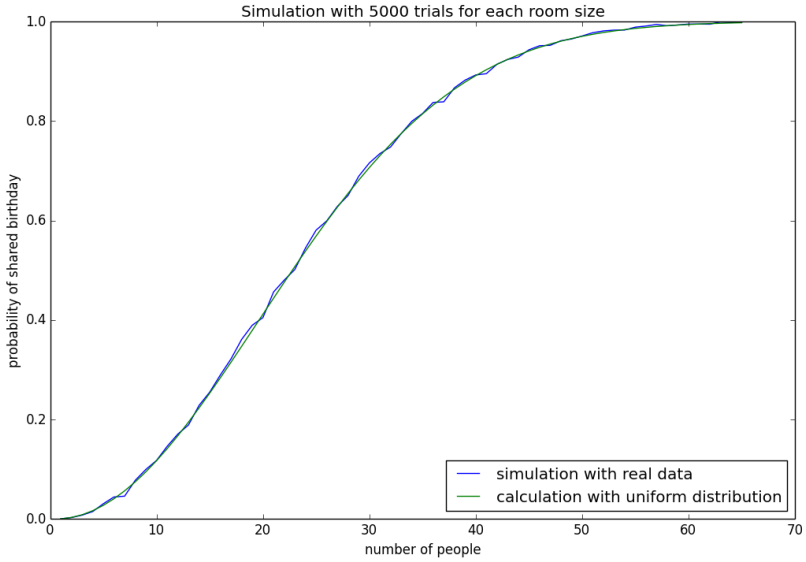
Some takeaways–there is a lot of nonuniformity in the distribution, especially for a single year! There does seem to be a seasonal uptick in births around mid-September, and of course February 29 shows up as an outlier in the plot for a four-year period. What is most remarkable, in my views, is what shows up very strongly as several parallel series, especially in the data for a single year, the large

uptick around mid-December, and the fact that there are relatively few birthdays on Christmas Eve, Christmas Day, or New Year's Day.

The parallel series are tied to the day of the week (this gets blurred in the data for a 4-year period, and would be completely erased in a 7-year sample): Many birth dates are planned, either through scheduled Caesarean sections or scheduled dates to induce labor. The hospital staff doesn't like to do this on weekends! (Nor, apparently, on Mondays.) The same explanation accounts for the smaller number of births on Christmas and New Year's. (The mid-September rise may have a more 'natural' explanation.)

As the final plot shows, in spite of all this non-uniformity, when the question is the probability of a coincidental birthday, our simple theoretical model is stunningly accurate!


Distribution of US births in 2000-2003
proportion of births
day (1=Jan 1,366=Dec 31)
0.005
0.004
0.003
0.002
0.001
0.000
−0.001
−50
0
50
100
150
200
250
300
350
400


Simulation with 5000 trials for each room size
probability of shared birthday
number of people
simulation with real data
calculation with uniform distribution
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70