AMT Homework 7 Supplement ANSWERS
Dr. Osborne 2014

The purpose of this supplement is to illustrate some of the techniques used to handle physical situations in which the outcome is assigned a given probability.

1. This problem concerns the binomial distribution, discussed in class. Note that the letter *n* has replaced the letter *k* in this problem.

   (a) Determine the value of $\sum_{n=0}^{N}\binom{N}{n}n^3\,p^n(1-p)^{N-n}$.

   Repeated differentiation gives $\langle n^3\rangle = pN\left[p^2(N-1)(N-2)+3p(N-1)+1\right]$.

   (b) Determine the expectation value of the *product* of the number of heads and the number of tails. Does this equal the product of the expectation values? Explain why or why not.

   $\langle n(N-n)\rangle = N(N-1)p(1-p) \neq \langle n\rangle\langle N-n\rangle = N^2p(1-p)$. Difference of $-Np(1-p)$ is attributed to correlations; if *n* is *larger*, then $N-n$ must be *smaller*.

   (c) The relative error expected in assuming that the number of positive outcomes in any given trial is the expectation value $\overline{n}$ is given by $\sigma_n/\overline{n}$. Calculate this quantity for the binomial distribution. What happens as *N* increases?

   $\frac{\sigma_n}{\overline{n}} = \sqrt{\frac{1-p}{Np}}$. It gets smaller.

   (d) An unfair coin has a 75% probability that it will come up heads. This coin is tossed 1000 times. Determine $\overline{n}$ and σ_n in this situation.

   $\overline{n}\pm\sigma_n = 750\pm13.69$.

   (e) How many tosses of the coin in part (d) would you have to throw in order for the relative error of part (c) to be less than 1%?

   Approximately 3333.

   (f) Determine the probability that total number of heads thrown in part (d) lies within one standard deviation of the average. How does this compare to the continuous approximation?

   The number lying between 750 ± 14 heads is given by $\sum_{n=736}^{764}\binom{1000}{n}\left(\frac{3}{4}\right)^n\left(\frac{1}{4}\right)^{1000-n} \approx 0.710405$. The continuous value is 0.6935, a bit low owing to the continuous approximation itself as well as the fact that *p* does not lie particularly close to the middle. The continuous result for *exactly* one standard deviation is 0.6827, farther off owing to our rounding of 13.69.

(g) A gambler places bets that the number of heads attained in a run of 1000 flips lies between 760 and 770. What is the probability that he will win if he starts with a 'clean slate', no coins yet flipped.
17.876%. Analysis is similar to that in the last part.

(h) Suppose that the coin in part (g) has already been thrown 500 times, resulting in exactly 400 heads. What is the new probability that the gambler will win the whole run?
We have reduced the bet to between 360 and 370 in 500 flips, given by 26.256%. His chances are better because enough heads have already come up to partly offset the small amount his betting region lies from the mean. In fact, a bit too many heads have already come up and the gambler is highly likely to come up with *more than* 770 heads at the end of the day.

2. Suppose that there are n people in a room, randomly distributed according to birthday. Determine the probability of each of the following, and produce a graph of each of them together as well as a graph of their sum for n varying from 0 to 100. Ignore leap years, and assume that the probability that any given person has a given birth date is 1/365.

(a) All birthdays different.
(b) One pair.
(c) Two pair.
(d) Three pair.
(e) One triple.
(f) Two triples.
(g) A full house (triple and pair)
(h) A triple and two pairs.

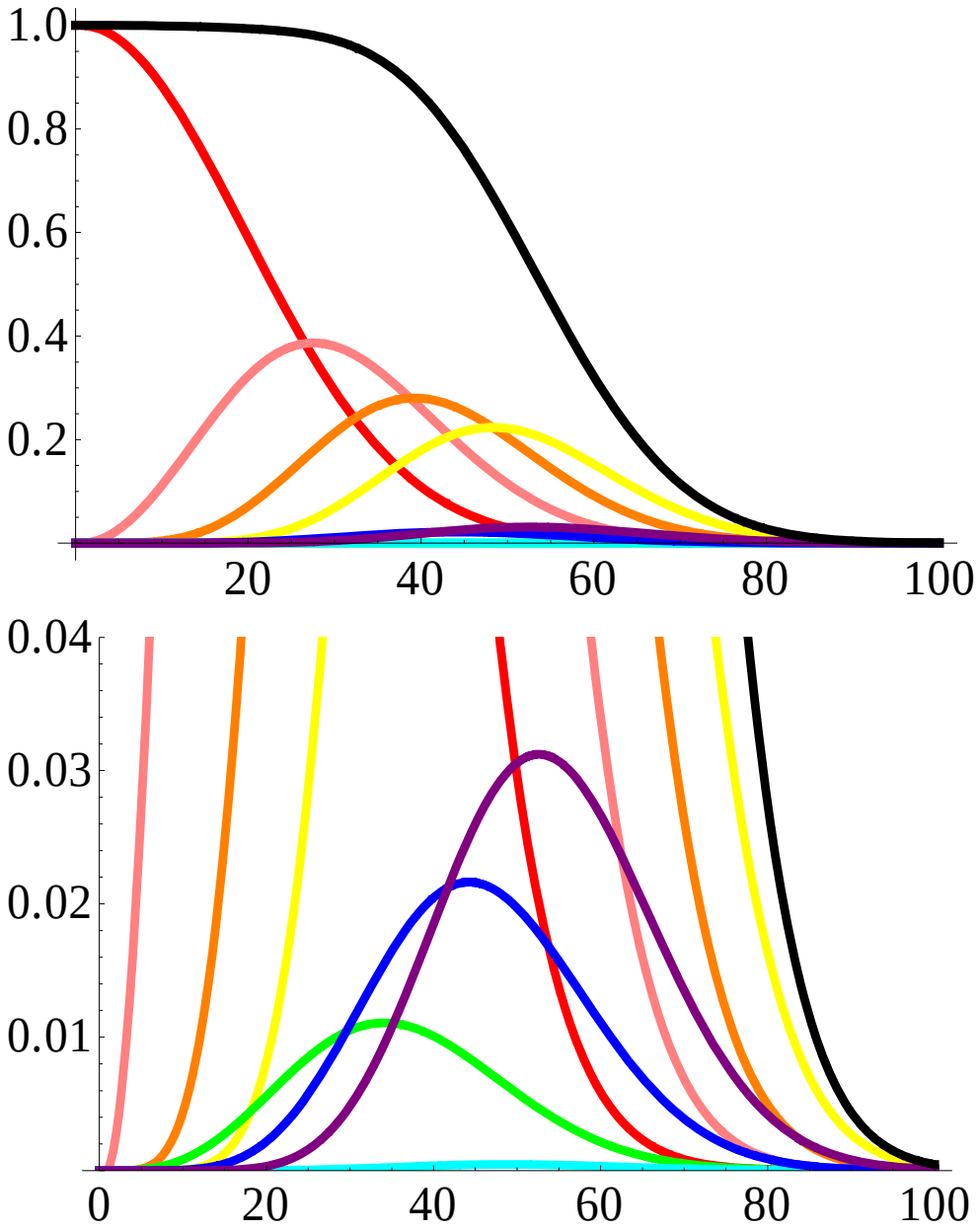
Write a sentence or two explaining which of these options dominates at various values of n. Explain what happens as n grows.

(a) All different is $\underbrace{\binom{365}{n}}_{\substack{\text{ways to}\\ \text{choose}\\ n \text{ days}}} \underbrace{n!}_{\substack{\text{ways to}\\ \text{distribute}\\ n \text{ days}}} \underbrace{\left(\frac{1}{365}\right)^n}_{\substack{\text{probability that}\\ \text{each person has}\\ \text{the birthday we want.}}}$.

(h) A triple and two pairs is

$$\underbrace{\binom{365}{2 \quad 1 \quad n-7 \quad 369-n}}_{\substack{\text{ways to choose two days for the double,}\\ \text{one for the triple, } n-7 \text{ for single birthdays,}\\ \text{and the rest of the days for no birthday.}}} \underbrace{\binom{n}{2 \quad 2 \quad 3 \quad n-7}}_{\substack{\text{ways to distribute the people}\\ \text{among the days, two on first}\\ \text{double, two on second, etc.}}} \underbrace{(n-7)!}_{\substack{\text{ways to}\\ \text{distribute}\\ \text{the single}\\ \text{birthdays}}} \underbrace{\left(\frac{1}{365}\right)^n}_{\substack{\text{probability that}\\ \text{each person has}\\ \text{the birthday we want.}}} .$$

The graph, including all of them, is given below. The black line curve is the sum, and the parts are given in 'rainbow' order. The triples are all very small, so an additional plot zooming in on those is given separately. Note that the probability of a single triple is smaller at its peak than that of a triple and a double, and that is in turn exceeded by a triple and two doubles as the number of people increases. This is entropy in action, as there are more ways to distribute these birthday configurations – once there are enough people.



3. Suppose that five different measurements of a given length are done, leading to the values 3.213 m, 3.219 m, 3.217 m, 3.221 m, and 3.215 m.

   (a) Use this data to estimate the mean and standard deviation of the underlying probability distribution governing the measurements.
   $\bar{x} \approx 3.217$; $\sigma \approx 0.00316$.

   (b) Using these approximations, determine the probability that the next measurement will yield a value larger than 3.25 m.
   3.25 m represents 10.4355 standard deviations above the mean, so the probability is given by $\frac{1}{\sqrt{2\pi}} \int_{10.4355}^{\infty} e^{-u^2/2} du = 8.5356 \times 10^{-26}$. Highly unlikely.

(c) Another measurement gave 3.258 m.
The new estimates of the mean and standard deviation are $\bar{x} \approx 3.224$; $\sigma \approx 0.017$. With these estimates, 3.25 m lies only 1.5294 standard deviations away from the mean… this modifies the probability to $\frac{1}{\sqrt{2\pi}} \int_{1.5294}^{\infty} e^{-u^2/2} du = 0.0630827$.

(d) Explain what part (c) illustrates about the nature of our *approximation* to the mean and standard deviation of the underlying distribution. How is this similar to the examples involving the 'hapless gambler' in the notes and in class?

We have no way of knowing what the next measurement will *actually* be. We can only approximate the exact results from the data we already *have*. *Any* of the data points we take can be wild statistical fluctuations that are totally irrelevant to the underlying distribution. We have no way of knowing, but the more data points we take, the more likely it is that our approximations are accurate.

4. (a) Show that $\sum_{n_1, n_2, \ldots, n_6 = 0}^{N} P_6^{(N)}(n_1, n_2, \ldots, n_6) = 1$, so the multinomial distribution is properly normalized. The sums take place independently over each n_i, but are subject to the constraint $\sum_{i=1}^{6} n_i = N$. This can be taken into account explicitly by simply writing $n_6 = N - \sum_{i=1}^{5} n_i$ and foregoing the sum over n_6 if you like.

$$\sum_{n_1, n_2, \cdots, n_6}^{N} \frac{N!}{n_1! n_2! \cdots n_5! n_6!} p_1^{n_1} p_2^{n_2} \cdots p_6^{n_6} = \sum_{n_1=0}^{N} \binom{N}{n_1} p_1^{n_1} \sum_{n_2=0}^{N-n_1} \binom{N-n_1}{n_2} p_2^{n_2} \cdots \times \sum_{n_5=0}^{N-n_1-\cdots-n_4} \binom{N-n_1-\cdots-n_4}{n_5} p_5^{n_5} p_6^{N-n_1-\cdots-n_5}$$

$$= \sum_{n_1=0}^{N} \binom{N}{n_1} p_1^{n_1} \sum_{n_2=0}^{N-n_1} \binom{N-n_1}{n_2} p_2^{n_2} \cdots \sum_{n_4=0}^{N-n_1-n_2-n_3} \binom{N-n_1-n_2-n_3}{n_4} p_4^{n_4} (p_5 + p_6)^{N-n_1-n_2-n_3-n_4}$$

$$= (p_1 + p_2 + \cdots + p_6)^N = 1$$

This process looks *far* more complicated than it actually is. Each of the sums basically collapses into its own binomial sum, and the collection of sums eventually becomes the expansion of $(p_1 + p_2 + \cdots + p_6)^N$, which is 1.

(b) Determine $\langle n_1 \rangle$ over this distribution by using a technique similar to that illustrated in problem 2. Note that the binomial expansion admits the generalization

$$(x + y + z)^N = \sum_{n_1, n_2 = 0}^{N} \frac{N!}{n_1!\, n_2!\, (N - n_1 - n_2)!} x^{n_1} y^{n_2} z^{N-n_1-n_2}, \text{ etc.}$$

$\langle n_1 \rangle = p_1 N$.

(c) Determine $\langle n_1 n_2 \rangle$ for this distribution. Is this equal to $\langle n_1 \rangle \langle n_2 \rangle$? Should it be? Why or why not?

$\langle n_1 n_2 \rangle = N(N-1) p_1 p_2$. This is smaller than $\langle n_1 \rangle \langle n_2 \rangle = N^2 p_1 p_2$ because n_1 and n_2 are correlated with each other. If n_1 is larger, then n_2 must be correspondingly smaller on average because the sum of all of the *n*'s is fixed.

Consider the specific case of a six-sided fair die. The probability of getting any outcome from 1 through 6 in any specific roll is $1/6$.

(d) Determine the expectation value of the product of the number of '2's, the number of '3's, and the number of '6's attained in a run of *N* rolls of a fair die. Why doesn't this equal $(N/6)^3$? What is the difference between your result and this naïve expectation? Explain the reason why the naïve expectation is incorrect.

$\langle n_2 n_3 n_6 \rangle = N(N-1)(N-2)/6^3$; $\langle n_2 n_3 n_6 \rangle - \langle n_2 \rangle \langle n_3 \rangle \langle n_6 \rangle = -N(3N-2)/6^3$.

These quantities are different because of the correlation. One easy way to see this is to consider what happens if there are only *two* rolls. In that case, the expectation value of the product is clearly zero. The product of the expectation values, on the other hand, is clearly *not*.

(e) Given *N* rolls of the die, the total amount rolled is given by

$S = n_1 + 2n_2 + 3n_3 + 4n_4 + 5n_5 + 6n_6$, with $n_6 = N - \sum_{i=1}^{5} n_i$. Determine the expectation value of *S* and the standard deviation from this value. Note that the fact that the die is fair implies that the value of the index does not matter when calculating probabilities… $\langle n_1 n_2 \rangle = \langle n_3 n_5 \rangle$, etc. Does this mean that $\langle n_1 n_2 \rangle = \langle n_1^2 \rangle$?

In order to do this more generally, consider the sum $S = \sum_{k=1}^{m} s_k n_k$. In our case, $s_k = k$ and *m* = 6. $\langle S \rangle = \sum_{k=1}^{m} s_k \langle n_k \rangle = N \sum_{k=1}^{m} s_k p_k = \frac{N}{6} \sum_{k=1}^{6} k = 7N/2$. As for the standard deviation, we have

$$\sigma_s^2 = \left\langle \left(S - \langle S \rangle \right)^2 \right\rangle = \left\langle \left(\sum_{k=1}^{m} s_k \left(n_k - \langle n_k \rangle \right) \right)^2 \right\rangle$$

$$= \sum_{k=1}^{m} \sum_{j=1}^{m} s_k s_j \left\langle \left(n_k - \langle n_k \rangle \right) \left(n_j - \langle n_j \rangle \right) \right\rangle$$

.

The remaining expectation value separates into two pieces, one of which is trivial while the other can be done by differentiating… I will present the result in index notation:

$$\left\langle \left(n_k - \langle n_k \rangle\right)\left(n_j - \langle n_j \rangle\right)\right\rangle = \langle n_k n_j \rangle - \langle n_k \rangle \langle n_j \rangle$$
$$= N\left(\delta_{jk} p_j + (N-1) p_j p_k\right) - N^2 p_j p_k = N\left(\delta_{jk} p_j - p_j p_k\right) \quad ; \text{ NO sum.}$$

Einstein's summation convention has been *suspended* in this expression; the index *j* is *not* summed over. This expression leads immediately to

$$\sigma_s^2 = N\sum_{k=1}^{m} s_k^2 p_k - N\left(\sum_{k=1}^{m} s_k p_k\right)^2$$
$$= N\left[\frac{6\cdot 7\cdot 13}{6\cdot 6} - \left(\frac{6\cdot 7}{2\cdot 6}\right)^2\right] = \frac{7N}{24}(52-42) = \frac{35}{12}N \; .$$

(f) Suppose that the fair die is rolled 3 times. The sum of all rolls therefore varies from 3 to 18. What is the expectation value of *S* and its standard deviation in this case? Determine the probability that the three rolls give a total of 7, 16, and 17. Remember that you have to consider all of the different ways in which these numbers can be obtained from rolls… for example, the probability of obtaining a total of 15 must include rolls of (3, 6, 6) (there are $\frac{3!}{2!1!1!0!0!0!} = \binom{3}{2} = 3$ ways to do this, counting the location of the 3's and 6's), (4, 5, 6) (6 ways), and (5, 5, 5) (1 way). Therefore, the probability of rolling 15 is $(3+6+1)/6^3 = 0.0463$. $\bar{S} \pm \sigma_s = 10.5 \pm 2.958$. Total of 7 is represented by (5,1,1) (3 ways), (4,2,1) (6 ways), (3,3,1) (3 ways), and (3,2,2) (3 ways), so the probability is 6.94%. 16 has a probability of 2.78% and 17 is 1.39%. All of the probabilities are given by $(1 \quad 3 \quad 6 \quad 10 \quad 15 \quad 21 \quad 25 \quad 27 \quad 27 \quad 25 \quad 21 \quad 15 \quad 10 \quad 6 \quad 3 \quad 1)/6^3$, and you can easily see that they sum to 1.

(g) Explain why the process of rolling three dies at the same time is identical to that considered in part (e), as long as the interaction between the dies is ignored. What is the expectation value and standard deviation of the total obtained when rolling two dies simultaneously? Determine the probability that rolling two dies will yield a total of 7, the 'magic' craps number.

If we ignore the interaction between the dies, then three dies thrown simultaneously are equivalent to three dies thrown separately, almost by definition. In this case, $\bar{S} \pm \sigma_s = 7 \pm 2.415$. 'Magic' 7 boasts the maximum probability of 16.7%. This can be thought of as representing the fact that a 7 can be obtained by *any* starting number. If the first number is a '3', then the probability that we end up with 7 is simply the probability that the second die comes up '4'.

5. A room contains 7 people, selected at random. You are interested in determining the probability of various distributions of the day of the week on which the people were born. Assume that all days of the week are equally likely.

(a) Determine the probability that all people were born on different days of the week.

This is given by $\binom{7}{1\ 1\ 1\ 1\ 1\ 1\ 1}\left(\frac{1}{7}\right)^7 = 0.612\%$.

(b) Determine the probability that exactly two people were born on the same day of the week.

As before, we have $\binom{7}{2\ 1\ 1\ 1\ 1\ 1\ 0}\binom{7}{1}\binom{6}{1}\left(\frac{1}{7}\right)^7 = 12.85\%$. The new factors of 7 choose 1 and 6 choose 1 are associated with us having to *choose* the day that the pair shares as well as the day left out. The large multinomial distribution factor is associated with us choosing the *people*… 2 born on one day, one each on 5 other days. We can frame this second factor associated with choosing the *days* in the same way if we like:

$$\underbrace{\binom{7}{2\ 1\ 1\ 1\ 1\ 1\ 0}}_{\text{2 people on one day, one person on others}}\ \underbrace{\binom{7}{1\ 1\ 5}}_{\substack{\text{one day with a pair,}\\ \text{one day left out, and}\\ \text{5 days with one person.}}}\ \left(\frac{1}{7}\right)^7 = 12.85\%.$$

This will be useful, so I thought I'd flesh it out…

(c) Determine the probability that exactly two *pairs* of people were born on the same day of the week.

This is $\underbrace{\binom{7}{2\ 2\ 1\ 1\ 1\ 0\ 0}}_{\substack{\text{two people on one day, two people on}\\ \text{another day, one person each on three}\\ \text{days, and two days left out.}}}\ \underbrace{\binom{7}{2\ 2\ 3}}_{\substack{\text{two days with two}\\ \text{people, two days left out,}\\ \text{and three days with one person.}}}\ \left(\frac{1}{7}\right)^7 = 32.13\%$.

(d) Determine the probability that exactly three people were born on the same day of the week and none of the other people share a day of the week on which they were born.

This is $\binom{7}{3\ 1\ 1\ 1\ 1\ 0\ 0}\binom{7}{1\ 4\ 2}\left(\frac{1}{7}\right)^7 = 10.71\%$.

(e) Show that the sum of the results for parts (a) – (d) is less than 1, and indicate what distributions the remaining probability consists of. Pick one of these, and show that its probability is smaller than the difference between your sum and 1.
One example is the probability that three pairs of people have the same birthday, $\binom{7}{2\ 2\ 2\ 1\ 0\ 0\ 0}\binom{7}{3\ 1\ 3}\left(\frac{1}{7}\right)^7 = 10.71\%$, another has both a pair and a triplet, $\binom{7}{2\ 3\ 1\ 1\ 0\ 0\ 0}\binom{7}{1\ 1\ 2\ 3}\left(\frac{1}{7}\right)^7 = 21.42\%$, and yet another has a triplet and *two* pairs, $\binom{7}{2\ 2\ 3\ 0\ 0\ 0\ 0}\binom{7}{2\ 1\ 4}\left(\frac{1}{7}\right)^7 = 2.677\%$. The sum of parts (a) – (d) gives 56.302%. Adding the three possibilities here brings us to 91.11%. Other prominent outcomes are two triples, at 1.785%, one quadruple, 3.57%, and one quadruple and one pair, 2.677%. The other 5 possibilities all have probabilities less than one percent.

(f) Determine the expectation value of the product of the number of people born on Tuesday and the number of people born on Saturday, as well the standard deviation of this quantity. Why isn't this expectation value just equal to 1? Explain.
The expectation value is given by $\langle n_t n_s \rangle = N(N-1)p_t p_s = 6/7$. It is not equal to 1 because these two are correlated; the total number of people is fixed. The standard deviation is more complicated, but can be obtained by repeated differentiation. The result is $\langle n_t^2 n_s^2 \rangle = N(N-1)p_t p_s\left[1+(N-2)(p_t+p_s)+(N-2)(N-3)p_t p_s\right]$, so $\sigma_{ts}^2 = N(N-1)p_t p_s\left[1+(N-2)(p_t+p_s)-2(2N-3)p_t p_s\right] = 97\cdot 6/7^3$. The value of the product is therefore predicted to be 0.857 ± 1.303. The fact that the standard deviation is larger than 1 is an indication that probabilities are largest for $n = 0$ or 1. The probability of having two people with birthdays on a *given* day is only $\binom{7}{2}\left(\frac{1}{7}\right)^2\left(\frac{6}{7}\right)^5 = 19.83\%$, while the probabilities of one and zero people born on a given day are given respectively by $\binom{7}{1}\left(\frac{1}{7}\right)\left(\frac{6}{7}\right)^6 = 39.657\%$ and $\binom{7}{0}\left(\frac{1}{7}\right)^0\left(\frac{6}{7}\right)^7 = 33.99\%$. This lopsided behavior of the distribution has led to our strange result, as zero is certainly more probable than 2. As the number of people increases, it will cease to be a problem.