Problem 1

(Story Proof) Define $\left\{ \begin{array}{c} n \\ k \end{array} \right\}$ as the number of ways to partition $\{1, 2, \ldots, n\}$ into k non-empty subsets, or the number of ways to have n students split up into k groups such that each group has at least one student. For example, $\left\{ \begin{array}{c} 4 \\ 2 \end{array} \right\} = 7$ because we have the following possibilities:

$$\begin{array}{ll} \bullet\{1\},\{2,3,4\} & \bullet\{1,2\},\{3,4\} \\ \bullet\{2\},\{1,3,4\} & \bullet\{1,3\},\{2,4\} \\ \bullet\{3\},\{1,2,4\} & \bullet\{1,4\},\{2,3\} \\ \bullet\{4\},\{1,2,3\} & \end{array}$$

Prove the following identities:
(a)

$$\left\{ \begin{array}{c} n+1 \\ k \end{array} \right\} = \left\{ \begin{array}{c} n \\ k-1 \end{array} \right\} + k \left\{ \begin{array}{c} n \\ k \end{array} \right\}.$$

Hint: I'm either in a group by myself or I'm not.
(b)

$$\sum_{j=k}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) \left\{ \begin{array}{c} j \\ k \end{array} \right\} = \left\{ \begin{array}{c} n+1 \\ k+1 \end{array} \right\}.$$

Hint: First decide how many people are not going to be in my group.

Solution:
(a) When n turns to $n+1$, there may be two possible case

1, the new item becomes a new subset which is independent to the subset of $\left\{ \begin{array}{c} n \\ k-1 \end{array} \right\}$, then in this case we have $\left\{ \begin{array}{c} n \\ k-1 \end{array} \right\}$ because the new item is fixed in one subset.

2, the new item belongs to one subset of $\left\{ \begin{array}{c} n \\ k \end{array} \right\}$. since item can join any any one of the existing k sunsets, in this case we have $k \left\{ \begin{array}{c} n \\ k \end{array} \right\}$ possibilities. In conclusion, we can prove that

$$\left\{ \begin{array}{c} n+1 \\ k \end{array} \right\} = \left\{ \begin{array}{c} n \\ k-1 \end{array} \right\} + k \left\{ \begin{array}{c} n \\ k \end{array} \right\} \tag{1}$$

(b) According to the identities in sub-problem(a), if we extend k to $k+1$, then we have:

$$\left\{ \begin{array}{c} n+1 \\ k+1 \end{array} \right\} = \left\{ \begin{array}{c} n \\ k \end{array} \right\} + (k+1) \left\{ \begin{array}{c} n \\ k+1 \end{array} \right\} \tag{2}$$

then the identity to be proved in sub-problem(b) is equivalent to

$$\sum_{j=k}^{n-1} \left(\begin{array}{c} n \\ j \end{array} \right) \left\{ \begin{array}{c} j \\ k \end{array} \right\} = (k+1) \left\{ \begin{array}{c} n \\ k+1 \end{array} \right\} \tag{3}$$

Then we try to observe the generation process of $\left\{ \begin{array}{c} n \\ k+1 \end{array} \right\}$. Firstly consider then situation that we have k subsets, obviously we need at least $j = k$ items to make sure that each subset contains at least one item. j should be at most $n-1$ because we require at least one item to fill the last($k+1$th) subset. For each j, obviously we have $\left(\begin{array}{c} n \\ k-1 \end{array} \right)$ choices. Noting that in this process there may be repetitive situation because any one of the $k+1$ subsets can be seen as the last subset which is not considered in the potability discussion of $\left\{ \begin{array}{c} j \\ k \end{array} \right\}$, which means that $\sum_{j=k}^{n-1} \left(\begin{array}{c} n \\ j \end{array} \right) \left\{ \begin{array}{c} j \\ k \end{array} \right\}$ repeats $k+1$ times of $\left\{ \begin{array}{c} n \\ k+1 \end{array} \right\}$, therefore we can prove the identity in equ3, which is equivalent to the original identity in problem(b).

Problem 2

A *norepeatword* is a sequence of at least one (and possibly all) of the usual 26 letters a, b, c, .., z, with repetitions not allowed. For example, "course" is a norepeatword, but "statistics" is not. Order matters, e.g., "course" is not the same as "source". A norepeatword is chosen randomly, with all norepeatwords equally likely. Show that the probability that it uses all 26 letters is very close to $1/e$.

Solution:
Let n represents the length of the norepeatword, $N(n)$ repesents the number of norepeatword whose length is n. Obviously $N(n)$ is equal to the num of ordered sample with no replacement, which is equal to $\frac{K!}{(K-n)!}$, where $K = 26$. Thus the probability that the length of the selected norepeatword letter is 26 can be written as:

$$\frac{K!}{\sum_{j=1}^{K} \frac{K!}{(K-j)!}} = \frac{1}{\sum_{j=1}^{K} \frac{1}{(K-j)!}} = \frac{1}{\sum_{i=0}^{K-1} \frac{1}{i!}} \tag{4}$$

noting that the limitation of the numerator in above number is $1/e$ because the limitation of $\sum_{i=1}^{K} \frac{1}{i!}$ is $e-1$. Thus we can show that the probability that it uses all 26 letters is very close to $1/e$.

Problem 3

Given $n \geq 2$ numbers $(a_1, a_2, \ldots, a_n)$ with no repetitions, a bootstrap sample is a sequence $(x_1, x_2, \ldots, x_n)$ formed from the a_j 's by sampling with replacement with equal probabilities. Bootstrap samples arise in a widely used statistical method known as the bootstrap. For example, if $n = 2$ and $(a_1, a_2) = (3, 1)$, then the possible bootstrap samples are $(3, 3), (3, 1), (1, 3)$, and $(1, 1)$.
(a) How many possible bootstrap samples are there for $(a_1, \ldots, a_n)$?
(b) How many possible bootstrap samples are there for $(a_1, \ldots, a_n)$, if order does not matter (in the sense that it only matters how many times each a_j was chosen, not the order in which they were chosen)?
(c) One random bootstrap sample is chosen (by sampling from $a_1, \ldots, a_n$ with replacement, as described above). Show that not all unordered bootstrap samples (in the sense of (b)) are equally likely. Find an unordered bootstrap sample $\mathbf{b}_1$ that is as likely as possible, and an unordered bootstrap sample $\mathbf{b}_2$ that is as unlikely as possible. Let p_1 be the probability of getting $\mathbf{b}_1$ and p_2 be the probability of getting $\mathbf{b}_2$ (so p_i is the probability of getting the specific unordered bootstrap sample $\mathbf{b}_i$). What is p_1/p_2 ? What is the ratio of the probability of getting an unordered bootstrap sample whose probability is p_1 to the probability of getting an unordered sample whose probability is p_2 ?

Solution:
(a) It is obvious that every time we can choose any one of the n numbers, with n times, so there are n^n possible samples in total.
(b) According to Bose-Einstein Counting, the result is given by $\binom{n+n-1}{n} = \binom{2n-1}{n}$
(c) Since the order does not matter, so a bootstrap with all the elements different has the highest probability to be chosen, since there are many ordered permutations can be merged to it. On the contrary, bootstraps with all the elements the same have the lowest probability to be chosen. According to the statement, p_1 and p_2 can be respectively given by $p_1 = n!/n^n$ and $p_2 = 1/n^n$. Thus, we have $p_1/p_2 = n!$. Because there are n bootstraps which contains the identical number in it, so the probability of getting a bootstrap whose probability is p_1 to the probability of getting a bootstrap whose probability is p_2 is given by $P = n!/n = (n-1)!$.

Problem 4

(Geometric Probability) You get a stick and break it randomly into three pieces. What is the probability that you can make a triangle using such three pieces?

Solution:
It is denoted that the length of the three pieces are x, y, and $1-x-y$, respectively. It is obvious that we have $0 \leq x \leq 1$, $0 \leq y \leq 1$ and $0 \leq (1-x-y) \leq 1$. In order to make a triangle successfully, it is necessary that

- $x+y>1-x-y \Longrightarrow x+y>1/2$;
- $x+1-x-y>y \Longrightarrow y<1/2$;
- $y+1-x-y>x \Longrightarrow x<1/2$.

With the help of Figure 1, the probability is 1/4.

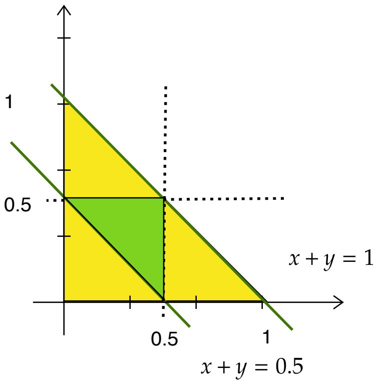

Figure 1: Problem 4.

Problem 5

In the birthday problem, we assumed that all 365 days of the year are equally likely (and excluded February 29). In reality, some days are slightly more likely as birthdays than others. For example, scientists have long struggled to understand why more babies are born 9 months after a holiday. Let $\mathbf{p}=(p_1,p_2,\ldots,p_{365})$ be the vector of birthday probabilities, with p_j the probability of being born on the j th day of the year (February 29 is still excluded, with no offense intended to Leap Dayers). The k th elementary symmetric polynomial in the variables $x_1,\ldots,x_n$ is defined by

$$e_k\left(x_1,\ldots,x_n\right)=\sum_{1\leq j_1<j_2<\cdots<j_k\leq n} x_{j_1}\ldots x_{j_k}.$$

This just says to add up all of the $\left(\begin{array}{c}n\\k\end{array}\right)$ terms we can get by choosing and multiplying k of the variables. For example, $e_1\left(x_1,x_2,x_3\right)=x_1+x_2+x_3$, $e_2\left(x_1,x_2,x_3\right)=x_1x_2+$ $x_1x_3+x_2x_3$, and $e_3\left(x_1,x_2,x_3\right)=x_1x_2x_3$
Now let $k\geq 2$ be the number of people.
(a) Find a simple expression for the probability that there is at least one birthday match, in terms of $\mathbf{p}$ and

Problem 5 continued on next page...

an elementary symmetric polynomial.
(b) Explain intuitively why it makes sense that P (at least one birthday match) is minimized when $p_j = \frac{1}{365}$ for all j, by considering simple and extreme cases.
(c) The famous arithmetic mean-geometric mean inequality says that for $x, y \geq 0$

$$\frac{x+y}{2} \geq \sqrt{xy}.$$

This inequality follows from adding $4xy$ to both sides of $x^2 - 2xy + y^2 = (x-y)^2 \geq 0$ Define $\mathbf{r} = (r_1, \ldots, r_{365})$ by $r_1 = r_2 = (p_1 + p_2)/2, r_j = p_j$ for $3 \leq j \leq 365$. Using the arithmetic mean-geometric mean bound and the fact, which you should verify, that

$$e_k(x_1, \ldots, x_n) = x_1 x_2 e_{k-2}(x_3, \ldots, x_n) + (x_1 + x_2) e_{k-1}(x_3, \ldots, x_n) + e_k(x_3, \ldots, x_n)$$

show that $P($ at least one birthday match $\mid \mathbf{p}) \geq P($ at least one birthday match $\mid \mathbf{r})$ with strict inequality if $\mathbf{p} \neq \mathbf{r}$, where the given $\mathbf{r}$ notation means that the birthday probabilities are given by $\mathbf{r}$. Using this, show that the value of $\mathbf{p}$ that minimizes the probability of at least one birthday match is given by $p_j = \frac{1}{365}$ for all j.

Solution:
(a) It is easier to consider the problem inversely, i.e., what is the probability that no person shares the same birthday. Since $e_k(\mathbf{p})$ multiplies and sums k **different** days, and there are $k!$ ways to map each condition to everyone, so the probability for no matching can be given by $k!e_k(\mathbf{p})$, thereby the probability for at least one birthday matching can be given by $1 - k!e_k(\mathbf{p})$.
(b) By considering the simple case where $k = 2$, we have

$$\begin{aligned} & P(\text{at least one birthday matches}) \\ =& 1 - 2e_2(\mathbf{p}) \\ =& (\sum_{i=1}^{365} p_i)^2 - 2 \sum_{1 \leq i < j \leq n} p_i p_j \\ =& \sum_{i=1}^{365} p_i^2 \\ \geq& 365 \cdot \left(\frac{\sum_{i=1}^{365} p_i}{365} \right)^2 \\ =& \frac{1}{365}, \end{aligned} \tag{5}$$

where the equation holds if and only if $\forall i, p_i = 1/365$.
(c) For each terms in the expansion of $e_k(x_1, \cdots, x_n)$, it either contains x_1 or not. For those terms not containing x_1, their sum can be written as $e_k(x_2, \cdots, x_n)$. For those terms containing x_1, by extracting the common factor (i.e., x_1), they can be written as $x_1 \cdot e_{k-1}(x_2, \cdots, x_n)$ (this is correct, because there are $\binom{n}{k}$ terms in $e_k(x_1, \cdots, x_n)$, $\binom{n-1}{k}$ terms in $e_k(x_2, \cdots, x_n)$, $\binom{n-1}{k-1}$ terms in $e_{k-1}(x_2, \cdots, x_n)$, and $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$). By applying the same steps to x_2, this conclusion holds.

$$\begin{aligned} & P(\text{at least one birthday match}|\mathbf{p}) \\ =& 1 - k!e_k(\mathbf{p}) \\ =& 1 - k!(p_1 p_2 e_{k-2}(p_3, \cdots, p_n) + (p_1 + p_2)e_{k-1}(p_3, \cdots, p_n) + e_k(p_3, \cdots, p_n)) \\ \geq& 1 - k! \left(\frac{(p_1 + p_2)^2}{4} e_{k-2}(p_3, \cdots, p_n) + (p_1 + p_2)e_{k-1}(p_3, \cdots, p_n) + e_k(p_3, \cdots, p_n) \right) \\ =& 1 - k!(r_1 r_2 e_{k-2}(r_3, \cdots, r_n) + (r_1 + r_2)e_{k-1}(r_3, \cdots, r_n) + e_k(r_3, \cdots, r_n)) \\ =& 1 - k!e_k(\mathbf{r}) \\ =& P(\text{at least one birthday match}|\mathbf{r}) \end{aligned} \tag{6}$$

Problem 5 continued on next page...

The proposition can be then proved by contradiction. Assume that $\mathbf{p} = (365^{-1}, \cdots, 365^{-1})$, and $\mathbf{p}' = (p'_1, \cdots, p'_n) \neq \mathbf{p}$ satisfies that $\mathbf{p}'$ minimizes the probability. Since $\mathbf{p}' \neq \mathbf{p}$, there are at least two elements, saying p'_i and p'_j, satisfy that $p'_i \neq p'_j$. Then, a corresponding $\mathbf{r}' = (r'_1, \cdots, r'_n)$ can be further given by $r'_i = r'_j = (p'_i + p'_j)/2$, $r'_k = p'_k (k \neq i, j)$. It is obvious that we have $P(\text{at least one birthday match}|\mathbf{p}') \geq P(\text{at least one birthday match}|\mathbf{r}')$, which contradicts with the assumption. Therefore, the probability is minimized only when $\mathbf{p} = (365^{-1}, \cdots, 365^{-1})$.

Problem 6

(Coupon Collection) If each box of a brand of crispy instant noodle contains a coupon, and there are 108 different types of coupons. Given $n \geq 200$, what is the probability that buying n boxes can collect all 108 types of coupons? You also need to plot a figure to show how such probability changes with the increasing value of n. When such probability is no less than 95%, what is the minimum number of n ?

Solution:

To sample an arbitrary type of coupons from all 108 types for n times, there are 108^n possibilities in total. In order to collect all the 108 types with n boxes, it is equivalent to find a division of n, where it is divided into 108 non-empty subsets and each type of coupon is placed into the corresponding subsets, whose result is given by $\left\{ n \atop 108 \right\}$. Since the order of these types does not matter, there are 108! permutations in total, so the final probability can be written as $P = 108! \left\{ n \atop 108 \right\} / 108^n$.

With the help of the formula of Stirling number, i.e., $\left\{ n \atop m \right\} = \frac{1}{m!} \sum_{k=0}^{m} (-1)^k \binom{m}{k} (m-k)^n = \sum_{k=0}^{m} (-1)^k \frac{(m-k)^n}{k!(m-k)!}$, the probability can be expanded by $P = \sum_{k=0}^{108} (-1)^k \frac{108!}{k!(108-k)!} \cdot \left(\frac{108-k}{108} \right)^n$.

With numerical experiments, the minimal number of boxes for $P \geq 0.95$ is 823, as shown in Figure 2.

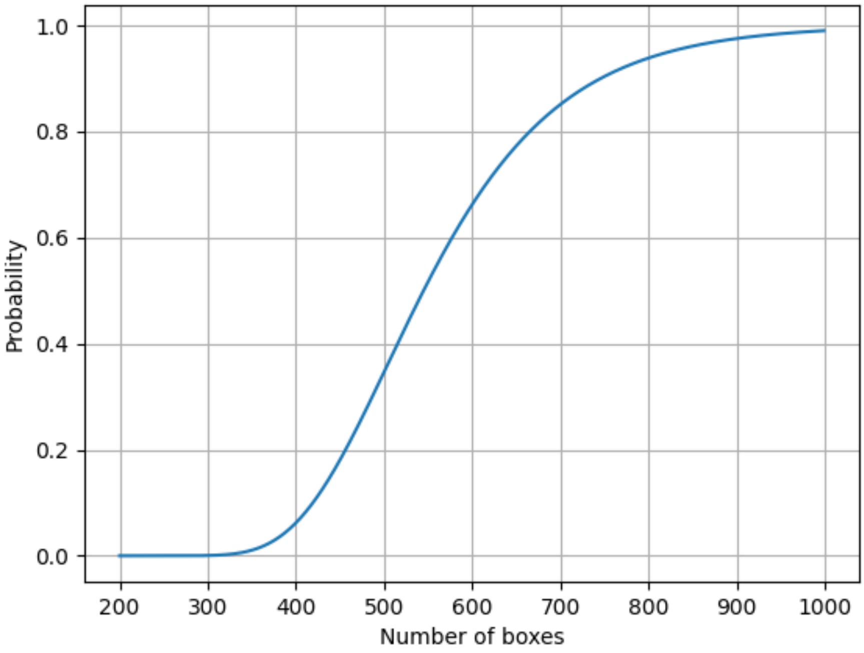

Figure 2: Problem 6.