2023Fall Probability & Statistics for EECS 2023/10/03

Homework 1

Professor: Ziyu Shao & Dingzhu Wen *Due: 2023/10/15 10:59pm*

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:

   - $\{1\}, \{2, 3, 4\}$
   - $\{2\}, \{1, 3, 4\}$
   - $\{3\}, \{1, 2, 4\}$
   - $\{4\}, \{1, 2, 3\}$
   - $\{1, 2\}, \{3, 4\}$
   - $\{1, 3\}, \{2, 4\}$
   - $\{1, 4\}, \{2, 3\}$

   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.

2. A *norepeatword* is a sequence of at least one (and possibly all) of the usual 26 letters $\mathrm{a}, \mathrm{b}, \mathrm{c}, \ldots, \mathrm{z}$, 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 $1/e$.

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?

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?

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 kth elementary symmetric polynomial in the variables $x_1, \ldots, x_n$ is defined by

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

This just says to add up all of the $\binom{n}{k}$ terms we can get by choosing and multiplying k of the variables. For example, $e_1(x_1, x_2, x_3) = x_1 + x_2 + x_3, e_2(x_1, x_2, x_3) = x_1x_2 + x_1x_3 + x_2x_3$, and $e_3(x_1, x_2, x_3) = x_1x_2x_3$ Now let $k \ge 2$ 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 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 $x, y \ge 0$

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

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

$$e_k(x_1, \ldots, x_n) = x_1x_2e_{k-2}(x_3, \ldots, x_n) + (x_1 + x_2)\, e_{k-1}(x_3, \ldots, x_n) + e_k(x_3, \ldots, x_n)$$

show that

$$P(\text{at least one birthday match} \mid \mathbf{p}) \geq P(\text{at least one birthday match} \mid \mathbf{r})$$

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.

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?