1. (Coupon Collector's Problem) A cereal company hosts a contest that requires you to collect all n types of coupons to win. Each cereal box you buy contains a coupon with each type equally probable. Let Y_n be the random variable representing the number of boxes you need to buy in order to collect all n coupons (clearly, $Y \geq n$ with probability 1).

   (a) Representing Y_n as a sum of appropriately chosen geometric random variables, compute $\mathbb{E}Y_n$ and $\text{Var}(Y_n)$.

   (b) Show that the pgf of Y_n is

   $$g_{Y_n}(t) = \frac{(n-1)!t^n}{(n-t)(n-2t)\cdots(n-(n-1)t)}$$

   Using the fact that $\mathbb{E}Y_n = \{\frac{d}{dt}\log g_{Y_n}(t)\}|_{t=1}$ and $\text{Var}(Y) = \{\frac{d}{dt}\log g_{Y_n}(t) + \frac{d^2}{dt^2}\log g_{Y_n}(t)\}|_{t=1}$, confirm your answer from part (a).

   (c) Define $Z_n = \frac{Y_n}{n\log(n)}$. Show that $Z_n \xrightarrow{\mathcal{P}} 1$.

2. (Sample Surveys) Suppose we have fixed observations $y_1, \cdots, y_N \geq 0$ and we hope to estimate $t = \sum_{i=1}^N y_i$. It's expensive to actually collect all observations, so we instead sample a subset $s \subseteq \{1, \cdots, N\}$ according to some sampling design. Denote $p_i = \mathbb{P}(i \in s)$.

   (a) Show that $\mathbb{E}|s| = \sum_{i=1}^N p_i$.

   (b) Let $\widehat{t} = \sum_{i\in s} \frac{y_i}{p_i}$. Show that $\mathbb{E}\widehat{t} = t$.

   (c) In our sampling design, let's say that we sample each index independently of others: $\mathbb{P}(i, j \in s) = p_i p_j$ for $i \neq j$. Subject to expected sample size n, find the values of p_i which minimize $\text{Var}(\widehat{t})$.

3. (Branching Processes)

   (a) Suppose spotted rabbits have a geometric offspring distribution with $p_k = (1-\theta)\theta^k$ for $k \geq 0$ and $\theta = \frac{2}{3}$. If the current world population of spotted rabbits if 10, what is the probability they will go extinct in 1 generation? What is the probability they will go extinct eventually?

   (b) Same two questions for mottled squirrels, whose offspring distribution is Poisson with mean 0.8, and current population 5000.

4. (Poisson Processes)

   (a) Let M_t, N_t be independent Poisson processes with rates 1 and 2, respectively. Find

      i. $\mathbb{P}(M_1 + N_1 = 4)$

ii. $\mathbb{P}(M_1 + N_2 = 4)$

iii. $\mathbb{P}(N_2 = 1 | M_1 + N_2 = 4)$

iv. $\mathbb{P}(N_1 = 1 | M_1 + N_2 = 4)$

(b) Let N_t be a rate 3.14159 Poisson process with arrival times $T_1 \leq T_2 \leq \cdots$. Find $\mathbb{P}(T_1 > 1 | T_{10} = 6)$.

(c) Let N_t be a rate λ Poisson process. Find the distribution of N_U, where $U \sim \text{Unif}(0, 1)$ and independent of N_t (you may leave your answer as an integral).

5. (Proof of weighted AM-GM inequality without using Jensen's inequality)

(a) Without using Taylor series, prove the inequality $e^{x-1} \geq x$ for all $x \in \mathbb{R}$ that Marcello has asked you to do so many times.

(b) For $a_i \geq 0$ and $p_i \geq 0$ such that $\sum_{i=1}^n p_i = 1$, use part (a) to show that

$$\exp\left(\left\{\sum_{i=1}^n p_i a_i\right\} - 1\right) \geq \prod_{i=1}^n a_i^{p_i}$$

(c) Using the substitution $a_i = x_i/\mu$, where $\mu = \sum_{i=1}^n p_i x_i$, prove that

$$\sum_{i=1}^n p_i x_i \geq \prod_{i=1}^n x_i^{p_i}$$

which is the inequality between the weighted arithmetic mean and weighted geometric mean (weighed AM-GM inequality).

6. (Limits of ratios) Let $X_1, X_2, \cdots \overset{\text{iid}}{\sim} \text{Unif}(0, 1)$. I claim that

$$\frac{\sum_{k=1}^n X_k^2}{\sum_{k=1}^n X_k} \overset{?}{\to} a$$

What is the strongest mode of convergence (in distribution, in probability, almost surely) that I can replace "?" with? What is a? Prove your claims.

7. (Making up data) Let $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$, and let $X_{(1)}, \cdots, X_{(n)}$ be the order statistics.

(a) Derive the pdf of $X_{(n)}$ (you cannot just state it). Show that $X_{(n)} \overset{\mathcal{P}}{\to} \theta$.

(b) State the pdf of the joint distribution $\mathbf{X}_{(n)} = (X_{(1)}, \cdots, X_{(n)})$.

(c) Suppose you were clumsy and lost observations $X_{(1)}, \cdots, X_{(n-1)}$, but still have $X_{(n)}$. You generate new random variables $\widetilde{X}_1, \cdots, \widetilde{X}_{n-1} | X_{(n)}$ such that, conditioned on $X_{(n)}$,

are conditionally independent and identically distributed as Unif$(0, X_{(n)})$, and let $\widetilde{X}_{(1)}, \cdots, \widetilde{X}_{(n-1)}$ be the order statistics of these new random variables. Show that

$$(\widetilde{X}_{(1)}, \cdots, \widetilde{X}_{(n-1)}, X_{(n)}) \stackrel{\mathcal{D}}{=} (X_{(1)}, \cdots, X_{(n-1)}, X_{(n)})$$

[Hint: Show $f(\widetilde{x}_{(1)}, \cdots, \widetilde{x}_{(n-1)}, x_{(n)}) = f(\widetilde{x}_{(1)}, \cdots, \widetilde{x}_{(n-1)} | x_{(n)}) f(x_{(n)})$ is equal to the pdf from part (b).]

8. (A bound for the expected maximum) Let $X_1, \cdots, X_n \stackrel{\text{iid}}{\sim} \text{Pois}(\lambda)$. Prove that

$$\mathbb{E} \max(X_1, \cdots, X_n) \leq \log(n) + \lambda(e-1)$$

[Hint: Somewhere along the way, you will need to make use of Jensen's inequality $e^{\mathbb{E}Y} \leq \mathbb{E}e^Y$.]

9. (Estimating location parameter in Cauchy) The pdf and cdf of a Cauchy(m, s) are

$$f_X(x) = \frac{1}{\pi s(1 + (\frac{x-m}{s})^2)}, \qquad F_X(x) = \frac{1}{\pi} \tan^{-1}\left(\frac{x-m}{s}\right) + \frac{1}{2}$$

Let $X_1, \cdots, X_n \stackrel{\text{iid}}{\sim} \text{Cauchy}(m, s)$. The sample median is defined as

$$\widetilde{X}_n = \begin{cases} X_{(\frac{n+1}{2})} & n \text{ odd} \\ \frac{X_{(n/2)} + X_{(n/2+1)}}{2} & n \text{ even} \end{cases}$$

Prove that $\widetilde{X}_n \stackrel{\mathcal{P}}{\to} m$.