Homework 6

Reading: Meester, Section 2.5, 3.1, 3.2, 4.2

Idea Journal: Remember to submit the idea journal entries after each class.

Problems/Exercises: All of these are due **Friday, March 8 by 5:00pm**.

1. Meester exercise 2.7.29. Some clarifying remarks about the experiment described: imagine that there are k bins, and there are n marbles that are sorted randomly into the bins, one at a time and independently. The probably that a given marble falls into the bin labeled i is p_i. The random variable N_i is the number of marbles that fall into bin i. So, $n_1 + n_2 + \cdots + n_k = n$, and the question is about the joint distribution of the random variables $N_1, \ldots, N_k$.

2. Let $p \in (0,1)$. Suppose the random variables $X_1, X_2, \ldots, X_n$ are independent and have the Bernoulli(p) distribution.

   (i) Compute the function
   $$\mu(\lambda) = \ln \mathbb{E}[e^{\lambda X_1}], \quad \lambda \in \mathbb{R}$$
   in terms of p and λ. In particular, what are $\mu(0)$ and $\mu'(0)$?

   (ii) Compute the function
   $$h(\lambda) = \ln \mathbb{E}[e^{\lambda(X_1 + \cdots + X_n)}].$$
   in terms of $\mu(\lambda)$ and n.

   (iii) Show that for any $\epsilon > 0$ and $\lambda > 0$,
   $$\mathbb{P}\left(X_1 + \cdots + X_n > n(p + \epsilon)\right) \leq e^{n(\mu(\lambda) - \lambda(\epsilon + p))}$$
   Hint: the form of this bound makes me think about Markov's inequality.

   (iv) **Extra, not required:** Choose $\lambda > 0$ to optimize the bound in (iii) (the quantity $(\mu(\lambda) - \lambda(\epsilon + p))$ will be negative for certain $\lambda > 0$), to derive a bound on
   $$\mathbb{P}\left(X_1 + \cdots + X_n > n(p + \epsilon)\right)$$
   which decays *exponentially* fast as $n \to \infty$. When $p = 1/2$, this should agree with the bound $e^{-\epsilon^2 n}$ from Theorem 0.1 in the notes *largenumber1.pdf*. In fact, the proof we gave for that Theorem is a particular case of the argument outlined above. This bound is a version of what is called Hoeffding's inequality. Another remark: the maximum value of $\lambda \mapsto (\lambda(\epsilon + p) - \mu(\lambda))$ turns out to be a quantity called the *relative entropy* between the Bernoulli($p+\epsilon$) distribution and the Bernoulli(p) distribution.

3. Consider random variables X and Y which have a joint distribution described as follows:

$$\begin{aligned} \mathbb{P}(X = 1,\ \ Y = 0) &= 1/4, \\ \mathbb{P}(X = -1,\ \ Y = 0) &= 1/4, \\ \mathbb{P}(X = 0,\ \ Y = 1) &= 1/4, \\ \mathbb{P}(X = 0,\ \ Y = -1) &= 1/4. \end{aligned}$$

   (i) What are the marginal distributions of X and of Y?

   (ii) Are X and Y independent?

   (iii) Compute the covariance $\mathrm{Cov}(X, Y)$.

4. Suppose there are N boxes, and each box has N marbles. In the k^{th} box, k of the N marbles are red, and the other $N - k$ of them are blue. Suppose that a box is chosen at random, but you do not know the fraction of red marbles it contains. Then, from that chosen box, you draw a marble repeatedly, with replacement after each draw. Let X_L be the number of red marbles after L draws from the box, so X_L is a random variable, taking a value in $\{0, 1, \ldots, L\}$. Let B be the index of the box chosen (a random variable taking values in $1, \ldots, N$).

   (i) What is the joint distribution of B and X_L? This will depend on parameters L and N.

   (ii) What is the conditional distribution of B given $X_L = j$?

   (iii) The conditional distribution from part (ii) is a distribution on integers $k = 1, \ldots, N$. What value of k maximizes this conditional probability? (i.e. maximizes $k \mapsto \mathbb{P}(B = k \mid X_L = j)$). This would be considered your "best guess" at the box index, given $X_\ell = j$. Hint: you can treat k like a continuous variable and use calculus to maximize....

5. Let S_n be a simple random walk on the integers, starting from $S_0 = 0$. A natural question is: does the walk reach some threshold b at or before time n? Let M_n denote the maximum of $S_0, S_1, \ldots, S_n$. Thus, $M_n \geq b$ holds if and only if $S_k = b$ for some $k \in \{1, 2, \ldots, n\}$. We want to know the distribution of M_n. Observe that $M_n \geq S_n$ must hold.

   Let $N_n^{max}(b)$ denote the number of paths for which $S_k \geq b$ at some time $k \in \{1, 2, \ldots, n\}$ (i.e. the paths that hit or exceed the threshold at some point). Let $N_n^+(b)$ be the number of paths for which $S_n > b$ (the paths that end up above b at time n). Assume that b is an even integer and n is odd, so that $\mathbb{P}(S_n = b) = 0$.

   (i) Explain carefully why $N_n^{max}(b) = 2 \cdot N_n^+(b)$. The idea here is similar to the reflection principle, Lemma 3.1.6.

   (ii) Using (i), write a formula for $\mathbb{P}(M_n \geq b)$ in terms of n and b (there will be a summation in your answer).