Please provide complete and well-written solutions to the following exercises.

Due February 3, 9AM, to be submitted in blackboard, under the Assignments tab.

# Homework 2

**Exercise 1.** You want to complete a set of 100 baseball cards. Cards are sold in packs of ten. Assume that each individual card in the pack has a uniformly random chance of being any element in the full set of 100 baseball cards. (In particular, there is a chance of getting identical cards in the same pack.) How many packs of cards should you buy in order to get a complete set of cards? That is, what is the expected number of cards you should buy in order to get a complete set of cards (rounded up to a multiple of ten)? (Hint: First, just forget about the packs of cards, and just think about buying one card at a time. Let $N$ be the number of cards you need to buy in order to get a full set of cards, so that $N$ is a random variable. More generally, for any $1 \leq i \leq 100$, let $N_i$ be the number of cards you need to buy such that you have exactly $i$ distinct cards in your collection (and before buying the last card, you only had $i-1$ distinct cards in your collection). Note that $N_1 = 1$. Define $N_0 = 0$. Then $N = N_{100} = \sum_{i=1}^{100}(N_i - N_{i-1})$. You are required to compute $\mathbf{E}N$. You should be able to compute $\mathbf{E}[N_i - N_{i-1}]$. This is the expected number of additional cards you need to buy after having already collected $i-1$ distinct cards, in order to see your $i^{th}$ new card.)

**Exercise 2.** You are trapped in a maze. Your starting point is a room with three doors. The first door will lead you to a corridor which lets you exit the maze after three hours of walking. The second door leads you through a corridor which puts you back to the starting point of the maze after seven hours of walking. The third door leads you through a corridor which puts you back to the starting point of the maze after nine hours of walking. Each time you are at the starting point, you choose one of the three doors with equal probability.

Let $X$ be the number of hours it takes for you to exit the maze. Let $Y$ be the number of the door that you initially choose.

- Compute $\mathbf{E}(X|Y=i)$ for each $i \in \{1,2,3\}$, in terms of $\mathbf{E}X$.
- Compute $\mathbf{E}X$.

**Exercise 3.** Let $X_1, \ldots, X_n$ be continuous random variables with joint PDF $f\colon \mathbf{R}^n \to [0,\infty)$. Assume that

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), \qquad \forall\, x_1,\ldots,x_n \in \mathbf{R}.$$

Show that $X_1, \ldots, X_n$ are independent.

**Exercise 4.** Let $\phi\colon \mathbf{R} \to \mathbf{R}$. We say that $\phi$ is **convex** if, for any $x, y \in \mathbf{R}$ and for any $t \in [0,1]$, we have

$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y).$$

Let $\phi\colon \mathbf{R} \to \mathbf{R}$. Show that $\phi$ is convex if and only if: for any $y \in \mathbf{R}$, there exists a constant $a$ and there exists a function $L\colon \mathbf{R} \to \mathbf{R}$ defined by $L(x) = a(x-y)+\phi(y)$, $x \in \mathbf{R}$, such that $L(y) = \phi(y)$ and such that $L(x) \leq \phi(x)$ for all $x \in \mathbf{R}$. (In the case that $\phi$ is differentiable, the latter condition says that $\phi$ lies above all of its tangent lines.)

(Hint: Suppose $\phi$ is convex. If $x$ is fixed and $y$ varies, show that $\frac{\phi(y)-\phi(x)}{y-x}$ increases as $y$ increases. Draw a picture. What slope $a$ should $L$ have at $x$?)

**Exercise 5** (Jensen's Inequality)**.** Let $X\colon \Omega \to [-\infty, \infty]$ be a random variable. Let $\phi\colon \mathbf{R} \to \mathbf{R}$ be convex. Assume that $\mathbf{E}\,|X| < \infty$ and $\mathbf{E}\,|\phi(X)| < \infty$. Then

$$\phi(\mathbf{E}X) \leq \mathbf{E}\phi(X).$$

(Hint: use Exercise 4 with $y := \mathbf{E}X$.) Deduce the **triangle inequality**:

$$|\mathbf{E}X| \leq \mathbf{E}\,|X|\,.$$

**Exercise 6** (Markov's Inequality)**.** Let $X\colon \Omega \to [-\infty, \infty]$ be a random variable. Then

$$\mathbf{P}(|X| \geq t) \leq \frac{\mathbf{E}\,|X|}{t}, \qquad \forall\, t > 0.$$

(Hint: multiply both sides by $t$ and use monotonicity of $\mathbf{E}$.)

**Exercise 7** (The Chernoff Bound)**.** Let $X$ be a random variable and let $r > 0$. Define $M_X(t) := \mathbf{E}e^{tX}$ for any $t \in \mathbf{R}$. Show that, for any $t > 0$,

$$\mathbf{P}(X > r) \leq e^{-tr} M_X(t).$$

Consequently, if $X_1, \ldots, X_n$ are independent random variables with the same CDF, and if $r, t > 0$,

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i > r\right) \leq e^{-trn}(M_{X_1}(t))^n.$$

For example, if $X_1, \ldots, X_n$ are independent Bernoulli random variables with parameter $0 < p < 1$, and if $r, t > 0$,

$$\mathbf{P}\left(\frac{X_1 + \cdots + X_n}{n} - p > r\right) \leq e^{-trn}(e^{-tp}[pe^t + (1-p)])^n.$$

And if we choose $t$ appropriately, then the quantity $\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - p) > r\right)$ becomes exponentially small as either $n$ or $r$ become large. That is, $\frac{1}{n}\sum_{i=1}^{n} X_i$ becomes very close to its mean. Importantly, the Chernoff bound is much stronger than either Markov's or Cheyshev's inequality, since they only respectively imply that

$$\mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - p\right| > r\right) \leq \frac{2p(1-p)}{r}, \quad \mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - p\right| > r\right) \leq \frac{p(1-p)}{nr^2}.$$

**Exercise 8** (Confidence Intervals)**.** Among 625 members of a bank chosen uniformly at random among all bank members, it was found that 25 had a savings account. Give an interval of the form $[a, b]$ where $0 \leq a, b \leq 625$ are integers, such that with about 95% certainty, if we sample 625 bank members independently and uniformly at random (from a very large bank membership), then the number of these people with savings accounts lies in the interval $[a, b]$. (Hint: if $Y$ is a standard Gaussian random variable, then $\mathbf{P}(-2 \leq Y \leq 2) \approx .95$.)

**Exercise 9** (Hypothesis Testing)**.** Suppose we run a casino, and we want to test whether or not a particular roulette wheel is biased. Let $p$ be the probability that red results from one spin of the roulette wheel. Using statistical terminology, "$p = 18/38$" is the null hypothesis, and "$p \neq 18/38$" is the alternative hypothesis. (On a standard roulette wheel, 18 of the 38 spaces are red.) For any $i \geq 1$, let $X_i = 1$ if the $i^{th}$ spin is red, and let $X_i = 0$ otherwise.

Let $\mu := \mathbf{E}X_1$ and let $\sigma := \sqrt{\text{var}(X_1)}$. If the null hypothesis is true, and if $Y$ is a standard Gaussian random variable

$$\lim_{n\to\infty} \mathbf{P}\left( \left| \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \right| \geq 2 \right) = \mathbf{P}(|Y| \geq 2) \approx .05.$$

To test the null hypothesis, we spin the wheel $n$ times. In our test, we reject the null hypothesis if $|X_1 + \cdots + X_n - n\mu| > 2\sigma\sqrt{n}$. Rejecting the null hypothesis when it is true is called a type $I$ error. In this test, we set the type $I$ error percentage to be 5%. (The type $I$ error percentage is closely related to the p-value.)

Suppose we spin the wheel $n = 3800$ times and we get red 1868 times. Is the wheel biased? That is, can we reject the null hypothesis with around 95% certainty?

**Exercise 10.** A community has $m > 0$ families. Each family has at least one child. The largest family has $k > 0$ children. For each $i \in \{1, \ldots, k\}$, there are $n_i$ families with $i$ children. So, $n_1 + \cdots + n_k = m$. Choose a child randomly in the following two ways.

**Method 1**. First, choose one of the families uniformly at random among all of the families. Then, in the chosen family, choose one of the children uniformly at random.

**Method 2**. Among all of the $n_1 + 2n_2 + 3n_3 + \cdots + kn_k$ children, choose one uniformly at random.

What is the probability that the chosen child is the first-born child in their family, if you use Method 1?

What is the probability that the chosen child is the first-born child in their family, if you use Method 2?

**Exercise 11.** Let $0 < p \leq \infty$. Show that, if $Y_1, Y_2, \ldots : \Omega \to \mathbf{R}$ converge to $Y : \Omega \to \mathbf{R}$ in $L_p$, then $Y_1, Y_2, \ldots$ converges to $Y$ in probability.

Then, show that the converse is false.

**Exercise 12.** Prove the following statement. Almost sure convergence does not imply convergence in $L_2$, and convergence in $L_2$ does not imply almost sure convergence. That is, find random variables that converge in $L_2$ but not almost surely. Then, find random variables that converge almost surely but not in $L_2$.

**Exercise 13.** Estimate the probability that 1000000 coin flips of fair coins will result in more than $501,000$ heads, using the Central Limit Theorem. (Some of the following integrals may be relevant: $\int_{-\infty}^{0} e^{-t^2/2}dt/\sqrt{2\pi} = 1/2$, $\int_{-\infty}^{1} e^{-t^2/2}dt/\sqrt{2\pi} \approx .8413$, $\int_{-\infty}^{2} e^{-t^2/2}dt/\sqrt{2\pi} \approx .9772$, $\int_{-\infty}^{3} e^{-t^2/2}dt/\sqrt{2\pi} \approx .9987$.) (Hint: use Bernoulli random variables.)

Casinos do these kinds of calculations to make sure they make money and that they do not go bankrupt. Financial institutions and insurance companies do similar calculations for similar reasons.