Homework 6

Due: Tuesday, November 2
Remember to show your work for each problem to receive full credit.

Problem 1 (15 points)

In each part below, we consider the result of $n$ independent tosses of a fair coin, for which $P($Heads$) = P($Tails$) = \frac{1}{2}$. Let $X$ be the number of heads in $n$ trials, and $Y = \frac{X}{n}$. Numerically evaluate and report all bounds and probabilities.

(a) For $n \in \{10, 100, 1000\}$, use Chebyshev’s inequality and Hoeffding’s bound, separately, to lower bound $P\left(\frac{2}{5} \leq Y \leq \frac{3}{5}\right)$.

(b) Using the binomial cumulative distribution function, evaluate the exact probabilities of the events in part (a) (see numpy.random.binomial and scipy.stats.binom). Compare the Chebyshev and Hoeffding’s bounds to the true probabilities, and discuss any trends.

(c) Using a normal approximation, obtain an asymptotic approximation of $P(4 \leq X \leq 6)$, $P(40 \leq X \leq 60)$ and $P(400 \leq X \leq 600)$ respectively for $n = 10$, $n = 100$, and $n = 1000$. Concretely, replace the binomial distribution with a Gaussian distribution with mean and variance exactly equal to the mean and variance of the binomial distribution, and compute Gaussian tail bounds with an implementation of the normal distribution CDF (see scipy.stats.norm). (To ensure that your approximation is accurate when approximating with the continuous distribution, please use the $\frac{1}{2}$ adjustment that was discussed in lecture.) Compare these bounds to the Chebyshev, Hoeffding and binomial tail bounds, and discuss any trends.
Problem 2 (25 points)

For parts (a) through (d) below, provide a numerical answer to 4 significant digits, as well as an exact answer in terms of the standard normal CDF $\Phi(\cdot)$, or the corresponding quantile function $\Phi^{-1}(\cdot)$. To do this, for example, you can write $\Phi(1) \approx 0.8413$.

(a) Cyrus’ Chocolate Chip Cookies Corporation bakes consumable cookies whose circumferences are normally distributed with parameters $\mu = 10$ cm, $\sigma = 0.2$ cm. The picky patrons will only consume cookies with circumferences 10.0 ± 0.3cm. What fraction of Cyrus’ cookies are likely to remain unconsumed?

(b) Connor offers to improve the quality control of Cyrus’ cookie production. Connor claims can reduce the value of $\sigma$. What value of $\sigma$ will ensure that no more than 1 percent of Cyrus’ cookies remain unconsumed?

(c) Let $X$ be a normal random variable with mean $\mu = 0$ and standard deviation $\sigma$. Use the standard normal table (Page 155 of your book) to compute the probabilities of the events $\{X \geq k\sigma\}$ and $\{|X| \leq k\sigma\}$ for $k = 1, 2, 3$.

(d) Take $\epsilon = \frac{1}{10}$ and $\sigma = \frac{1}{4}$ and calculate $\Pr(\sigma k - \epsilon \leq X \leq \sigma k + \epsilon)$ for $k = 0, 1, 2, 3$ and for $k = -1, -2, -3$.

(e) Write code to plot $\Pr(\sigma k - \epsilon \leq X \leq \sigma k + \epsilon)$ for $\epsilon = \frac{1}{10}$, $\sigma = \frac{1}{4}$ and $k = 0, \pm 1, \pm 2, \pm 3, \ldots, \pm 10$ (with $k$ on the $x$-axis, and $\Pr(\cdot)$ on the $y$-axis). Does this shape remind you of anything you have seen before? Set $\sigma = 2$ and compare the two plots.
Problem 3 (35 points)

The Law of Large Numbers states that averages of certain i.i.d. random variables converge in probability to the mean. We will now find a “pathological” counterexample. Let us define a continuous random variable $S$ over $(-\infty, \infty)$ as follows:

$$f_{S}(x) = \frac{1}{\pi(1 + x^2)}$$

(a) Why doesn’t the Law of Large Numbers apply to $S$? (Explain both in math and words why).

(b) Let $Z = \frac{X}{Y}$ where $X, Y$ are independent standard normal random variables. Take 1000 samples of $Z$, and plot those as a histogram against the PDF of $S$. Comment on how the two compare.

(c) Simulate an empirical mean $\hat{E}[Z] \approx \frac{1}{n} \sum_{i=1}^{n} Z_i$ for $n \in \{50, 500, 5000\}$ (we can simulate further if we please). You can do this by first sampling (and saving) 5000 samples into an array or list, then computing the average on the first 50 samples, the first 500 samples, and the first 5000 samples. You can repeat this process 10 times (i.e. run 10 trials). Report your results. Do the averages seem to converge?

(d) Given two independent variables $X, Y$ distributed “pathologically,” write down an equation for finding the density of the sum $C = X + Y$ in terms of the densities $f_X, f_Y$. Your equation should involve a single integral (over a single variable); you need not evaluate it (though you may do so if you choose).

Depending on your approach, you may find the following helpful:

$$\lim_{z \to \infty} \arctan(z) = \frac{\pi}{2} \quad \text{and} \quad \frac{d}{dz} \arctan(z) = \frac{1}{1 + z^2}$$

(e) We claim without proof (the integral gets a bit complicated) that our density evaluates to

$$f_C(c) = \frac{2}{\pi(c^2 + 4)}$$

We now wish to find the density of the average $A = \frac{C}{2}$. Prove and apply the following statement: if $X, Y$ are two continuous r.v.’s with densities $f_X, f_Y$ and $X = bY, b > 0$, then $f_X(x) = \frac{1}{b}f_Y(\frac{x}{b})$.

Hint: First solve for the CDF.

(f) What can we conclude about the average of i.i.d “pathological” r.v.’s $A_n = \frac{1}{n} \sum_{i=1}^{n} S_i$?
Problem 4 (25 points)

We define the covariance Cov\((X, Y)\) by

\[\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]\]

Assuming \(X, Y\) have positive variance, we define their correlation by

\[\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X] \text{Var}[Y]}}\]

We will prove that the correlation coefficient \(\rho\) is between \(-1\) and 1.

(a) Prove that

\[\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2 \text{Cov}(X, Y)\]

(b) Using part (a), show that

\[0 \leq \text{Var} \left[ \frac{X}{\sigma[X]} + \frac{Y}{\sigma[Y]} \right] = 2(1 + \rho(X, Y))\]

and that

\[0 \leq \text{Var} \left[ \frac{X}{\sigma[X]} - \frac{Y}{\sigma[Y]} \right] = 2(1 - \rho(X, Y))\]

(c) Show that \(-1 \leq \rho(X, Y) \leq 1\).