ECO 394D: Probability and Statistics

Homework 2

1. Let X be a (continuous) uniform random variable on $[0, 1]$.

   (A) Compute $P(X^2 \leq 0.25)$.

   (B) For any number a, compute $P(X^2 \leq a)$.

   (C) Find the PDF of the random variable $Y = X^2$.

   (D) Compute $E(Y)$ and $\text{var}(Y)$ directly from the PDF.

2. The expected value $\mu = E(X)$ is often interpreted as one's "best guess" for the value of the random variable X. In this problem, you will show the formal mathematical interpretation of what "best guess" means, in this context. Specifically, the claim is that, among all possible values a, the expected value $\mu = E(X)$ minimizes the quantity

$$E([X - a]^2) \, .$$

   Think of a as your guess for X, and $E([X-a]^2)$ as a measurement of the quality of your guess. The term $[X - a]^2$ is the squared difference between X and a; the difference is squared, so that positive and negative differences both count the same. The quantity to be minimized is the *mean squared error*, i.e., the expectation of the squared difference between X and a.

   Prove that $\mu = E(X)$ is the choice of a that minimizes the mean-squared error (i.e. the expression above). Here are some hints:

   (i.) Write the objective function as $E([X - \mu + \mu - a]^2)$. (You can always add and subtract the same constant without changing anything.) Expand the square, but keep the terms $X - \mu$ and $\mu - a$ intact.

   (ii.) Take expectations term by term, using properties of the expectation operator to simplify your expression. Examine the expression that results, and draw the appropriate conclusion.

3. Suppose that we have d independent standard normal random variables $Z_1, \ldots, Z_d$, where each $Z_i \sim N(0,1)$. We say that a continuous random variable X follows a chi-squared (χ^2) distribution with d degrees of freedom if

$$X \overset{D}{=} Z_1^2 + \cdots + Z_d^2 \, .$$

Remember that $\overset{D}{=}$ means "equal in distribution." We write this as $X \sim \chi^2_d$, and it is possible to show that the PDF of X is

$$f(x) = \frac{1}{2^{d/2}\Gamma(d/2)} \cdot x^{d/2-1} \cdot \exp(-x/2) \, ,$$

where $\Gamma(\cdot)$ is the gamma function, which generalizes the factorial function to continuous numbers (see Wikipedia). Note: you don't have to derive this PDF.

The chi-squared distribution plays an important role in statistical inference. For now, your job is simply to compute $E(X)$ and $\text{var}(X)$. Hint: use the relationship between the chi-squared and normal distribution, rather than the PDF directly. This is much easier.

4. Suppose that U is a random variable with a uniform distribution on $[0,1]$. Now suppose that f is the PDF of some continuous random variable of interest, that F is the corresponding CDF, and assume that F is invertible (so that the function F^{-1} exists and gives a unique value).

Show that the random variable $X = F^{-1}(U)$ has PDF $f(x)$—that is, that X has the desired PDF. Hint: use results on transformations of random variables.

This cute result allows you to simulate random variables with any known (invertible) CDF, assuming that you have a source of uniform random variables.

5. This simulation exercise will serve as a bridge between the probability and the statistics parts of the course. For now, you can take it as an illustration of the relationship between the expected value (i.e., the theoretical mean) of a random variable and the sample average constructed from a number of (independent) observations on that random variable.

   (A) Suppose that $X_N \sim \text{Binomial}(N, P)$ be the (random) number of successes in a sequence of N binary trials. Let $\hat{p}_N = X_N/N$ denote the proportion of observed successes. Calculate $E(\hat{p}_N)$ and $\text{sd}(\hat{p}_N)$.

   (B) Using your favorite software package, simulate 100 (or more) realizations of the random variable $\hat{p}_5$, assuming that the true $P = 0.5$. In other words, for each of the 100 realizations, you will simulate five coin flips to get one value of X_5, and then set $\hat{p}_5 = X_5/5$. This process is repeated 100 times (or more). Verify that the sample (i.e., Monte Carlo) mean and sample standard deviation of your simulated $\hat{p}_5$'s agree, at least approximately, with the theoretical mean and standard deviation computable from your result in (A).

(C) Now repeat the process in part (B) for $\hat{p}_{10}$, $\hat{p}_{25}$, $\hat{p}_{50}$, and $\hat{p}_{100}$.

(D) Make a graph that overlays two sets of points: (i.) the sample standard deviation of $\hat{p}_N$ versus N for the five values of N that you used in your simulations in Parts B-C; and (ii.) the corresponding theoretical standard deviations versus N, calculated from your result in Part A. Comment on the patterns you see in the graph.