ECON 5213: ADVANCED ECONOMETRICS

PROF. LE WANG

Problem Set #2

Part I: Analytical Questions

Question 1. [Distribution of A function of a Random Variable]. The following result is important in Econometrics and often used in Monte Carlo Simulation (we will discuss this later).

Suppose that I would like to construct a random variable, Y, that can take on six values $\{1, 2, 3, 4, 5, 6\}$ with the following distribution:

$$\Pr[Y = 1] = p_1$$
$$\Pr[Y = 2] = p_2$$
$$\Pr[Y = 3] = p_3$$
$$\Pr[Y = 4] = p_4$$
$$\Pr[Y = 5] = p_5$$
$$\Pr[Y = 6] = p_6$$

where $\sum p_i = 1$. Can you think of a function $f(\cdot)$ such that $Y = f(U)$ has the distribution above, where the random variable U is distributed from a standard uniform? (**For later: imagine** Y represent a dice, fair or not. We can construct a hypothetical dice from a standard uniform variable using a computer later.)

Question 2. [Distribution of A Function of a Random Variable]. The p-value for a specific case with two-tailed alternative hypothesis, $p(z)$, is defined as

$$p(z) = \Pr[|Z| \geq |z|]$$

where $Z \sim N(0, 1)$ meaning that Z is normally distributed. Show that $p(Z)$ is uniformly distributed. Note that you just need to apply the definition of the CDF and the property of a uniform distribution. It should be one-line proof. One thing would be particularly useful when you show the proof is: what is the relationship between $z_1$and z_2 when we know $p(z_1) \geq p(z_2)$? Note that this result is particularly useful for hypothesis testing and machine learning on False Discovery Rate and Better design algorithm to reduce the false discovery rate.

Question 3. [Expectation of A Function of a Random Variable]. In class, we define the expectation of a function, $Y = g(x)$, as follows

$$\mathbb{E}[Y] = \mathbb{E}[g(x)] = \begin{cases} \int_{-\infty}^{\infty} g(x)f(x)dx & \text{if } X \text{ is continuous} \\ \sum g(x_i) \cdot p(x_i) & \text{if } X \text{ is discrete} \end{cases}$$

Formally, the expectation of the function of a continuous variable is well defined if $\int_{-\infty}^{\infty} |g(x)| \, f(x)dx < \infty$. Suppose that $X_1, \ldots, X_k$ are continuous variables.

Show the following results hold.

(1) Using the definition of expectation, $\mathbb{E}[Y] = \mathbb{E}[g(x)] = \sum g(x_i) \cdot p(x_i) \qquad \text{if } X \text{ is discrete.}$ In class we show that this result holds for the continuous case. Try it yourself before looking at the answer.
(2) $\mathbb{E}[c] = c$, where c is a constant.
(3) $\mathbb{E}[c_1 X_1] = c_1 \mathbb{E}[X_i]$
(4) Let X be a discrete variable distributed with the PMF $p(x) \equiv \Pr[X = x]$. Let a, b_1, b_2 be some constants. Then,$\mathbb{E}[a + b_1 u_1(X) + b_2 u_2(X)] = a + b_1 \mathbb{E}[u_1(X)] + b_2 \mathbb{E}[u_2(X)]$. **Note that this is different from the continous case that we proved in class.**
(5) Using the definition above, suppose that X is continuously distributed with the PDF $f_X(x)$. Then, $F_X(x) = \Pr[X \leq x] = \mathbb{E}[\mathbb{I}(X \leq x)]$. I show one way to prove this in class. But here I would like you to take a slightly different approach.
    (a) Treat $\mathbb{I}[X \leq x]$ as a random variable, Y. Derive the distribution for Y, as we did in class.
    (b) Apply the definition of mathematical expecation to obtain $\mathbb{E}[Y]$.
(6) Now use a similar approach to show that $\Pr[X = x] = \mathbb{E}[\mathbb{I}[X = x]]$.

Question 3. [Expectation of the function of a random variable]. In machine learning, an important task is **classification**. Classification is about classifying an object into a particular group. For example, to identify whether or not an email is a spam or to predict whether or not someone will be elected into a public office. As you can immeditely recognize, this outcome of interest is actually a discrete variable, and **classification** is about predicting whether or not the outcome will be a particular value. Based on the distribution, one simplest possible classification algorithm is to classify or predict an outcome to be **the most likely outcome** (i.e., the value with the highest probability). This algorithm is intuitive, and can also be justified by minimizing the expected error.

Suppose that if you predict an outcome **incorrectly**, you receive an error of 1 and zero otherwise. What is the value that minimizing the expected error?

$$\min_{a} \quad \mathbb{E}[\mathbb{I}(X - a \neq 0)]$$

Question 4. [Distribution of the function of a random variable]. Let X be a continuously distributed random variable with the CDF $F_X(x)$ and the PDF $f_X(x)$, where $g(x)$ is a montone **decreasing** function. Then, Y is continuously distributed with the CDF the PDF

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{1}{g'(g^{-1}(y))} \right|$$

(1) As in class, draw a function to intuitively discuss the solution for the CDF $\Pr[Y \leq y]$ first
(2) Mathematically derive the solutions. Clearly state the assumptions or conclusions that you use to dervie the solutions.

Part 2: Computer Questions

Question 1. [Empirical CDF and PMF]. In Stata, type **webuse auto,clear** to read in the data used in class. Then calculate the following quantities for the variable called **rep78** (Repair Record 1978) using Stata.

(1) $F(3) = \Pr[\mathbf{rep78} \leq 3]$
(2) $\Pr[\mathbf{rep78} = 3]$.

Include your Stata code, output, and your answers when submitting your homework (via Canvas).