Stat414/614; Fall 2024; Worksheet 03; 20 Points; NAME:

This worksheet is based on Chapter 4 of Millard and Neerchal. Baiscs of the probability distributions, Binomial, Poisson and Normal distribution is expected to be covered in the pre-requisite courses. This model goes a bit further and covers additional probability distributions as well. Important additional skills to be learned here is handling pdfs and pmfs in R. The package `EnvStats` is an invaluable tool in performing basic tasks such as plotting, drawing samples repeatedly, comparing different distributions etc which form the foundation of modern statistical analysis.

1. Describe some experiment or observational study, specifying two characteristics (preferably one categorical and the other continuous) of interest. Define the population, and how you will take physical samples. What are the random variables of interest in your study? What kind of shape and characteristics do you think the probability distributions of these random variables will have? Why?

2. Suppose that a population is adequately described by a Gamma distribution with shape=3 and scale=2.

   a. Generate three samples, of size $n = 10, 100$ and 1000, respectively. For each sample, create a density histogram. Overlay the pdf of Gamma(shape=3,scale=2) on the created histogram. Comment on how increasing the sample size improves the fit between the sample histogram and its theoretical expectation.

   b. Compute the mean, Variance, Skew and CV of the sample with 1000 observations generated above and compare them to their true values given in equations (4.47) of Millard and Neerchal.

   c. Generate 100000 random numbers from this distribution and create a density histogram. Overlay the pdf of Gamma(shape=3,scale=2) on this histogram. Compute the mean, Variance, Skew and CV of this sample and compare them to their true values given in equations (4.47) of Millard and Neerchal.

3. Suppose the population is well-approximated by a Normal distribution $N(\mu = 20, \sigma = 6)$.

   (a) Use the appropriate R function and compute the quantiles (x_p) corresponding to $p =$ 1%,2.5%, 5%, 10% 50% 90% 95%, 97.5% and 99%. Compute the corresponding quantiles (z_p) for the standard normal distribution.

   (b) Plot x_p vs z_p. What do you see? What is the slope and intercept of this line?

4. Consider a random variable $X \sim Binomial(n, p)$. It is known that, for large n, the Binomial probabilities are well-approximated by a Normal random variable with mean $\mu_X = E(X)$ and variance $\sigma_X^2 = Var(X)$. Let us find out how large n should be for this approximation to be acceptably accurate.

   (a) For $p = 0.65$, and $n = 10$, compute the mean (μ_X) and standard deviations (σ_X.) Plot the CDF of this random variable and the plot the CDF of the approximating Normal random variable.

   (b) Repeat the above with larger values of n and comment on the quality of approximation.

   (c) Repeat the above exercise with a different value of p, for example $p = 0.1$.