Multivariate Statistical Analysis, Exercise 4, Fall 2011, Prof. S.K. Jeng

October 21, 2011 TA: H.C. Cheng

(Central Limit Theorem)

The random sample in this exercise is an $M \times N$ matrix $\mathbf{X}$, where the elements comes independently and identically (iid) from a negative binomial (r, p) distribution with probability mass function (pmf)

$$f(k) = \Pr(X = k) = \binom{k+r-1}{k} p^r (1-p)^k \quad \text{for } k = 0, 1, 2, \ldots,$$

with mean $\mu = \frac{r(1-p)}{p}$ and variance $\sigma^2 = \frac{r(1-p)}{p^2}$. For this exercise, $M = 1000, N = 30, r = 10, p = \frac{1}{2}$. Now make the transformation

$$y_m = \frac{\sum_{n=1}^{N} \mathbf{X}_{m,n}}{N} \quad m = 1, 2, \ldots M.$$

Please answer the following questions.

(a) Use the histogram method to plot the figure of approximated probability density function (pdf), cumulative density function (cdf) of $\sqrt{N}(y_m - \mu)/\sigma$ and the cdf of the standard normal distribution.

(b) For calculating the probability $\Pr(y \leq 11)$, we can derive it explicitly from the pmf of the negative binomial distribution; that is

$$\begin{aligned} \Pr(y \leq 11) &= \Pr(\sum_{n=1}^{30} x_n \leq 330) \\ &= \sum_{x=1}^{330} \binom{300+x-1}{x} \left(\frac{1}{2}\right)^{300} \left(\frac{1}{2}\right)^{x} \simeq 0.8916. \end{aligned}$$

However, this is difficult to calculation due to the factorials. Now use the central limit theorem to derive $\Pr(y \leq 11)$ analytically. Use the approximated cdf to derive it.

(c) Order the observations $y_m \quad m = 1, \ldots, M$ as $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$ and calculate its standard normal quantiles:

$$\Pr(Z \leq q_{(j)}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{q_{(j)}} e^{-z^2/2} dz = \frac{j - 1/2}{M}.$$

Plot $(q_{(j)}, y_{(j)})$ and calculate its correlation coefficient.