QBS 120 - Problem Set 3

Rob Frost

1. (Based on Rice 4.54) Let X, Y, and Z be independent RVs with variances $\sigma_X^2, \sigma_Y^2, \sigma_Z^2$. Let:

$$U = Z - X$$
$$V = Z - Y$$

Answer the following questions:

   (a) Find Cov(U, V) and $\rho_{U,V}$
   (b) If $U = Z + X$ and $V = Z + Y$, do the values of $Cov(U, V)$ and $\rho_{U,V}$ computed in part a) change? Explain.
   (c) How does $\rho_{U,V}$ change if σ_Z^2 is much larger than σ_X^2 or σ_Y^2?
   (d) How does $\rho_{U,V}$ change if σ_Z^2 is much smaller than σ_X^2 or σ_Y^2?
   (e) How do the answers for parts c) and d) relate to variable standardization?

2. (Based on Rice 4.64) Let X and Y be jointly distributed RVs with correlation $\rho_{X,Y}$; define the standardized random variables $\bar{X}$ and $\bar{Y}$ as:

$$\bar{X} = (X - E[X])/\sqrt{Var(X)}$$
$$\bar{Y} = (Y - E[Y])/\sqrt{Var(Y)}$$

Answer the following questions:

   (a) Show that $Cov(\bar{X}, \bar{Y}) = \rho_{X,Y}$.
   (b) Principal component analysis (PCA) is normally defined by the eigenvalue decomposition of the sample covariance matrix for multivariate data. Look at the R PCA function prcomp(). What is the impact of setting center=T and scale=T when calling prcomp()? When might this be desirable?

3. (Based on Rice 4.74) The number of offspring of an organism is a discrete random variable with mean μ and variance σ^2. Each of its offspring reproduces in the same manner. *Hint: use the Law of Total Expectation.*

   (a) Find the expected number of offspring in the third generation.
   (b) Find the variance of the number of offspring in the third generation.
   (c) Validate your answers to a) and b) via simulation with the number of offspring represented by a Poisson RV with $\lambda = 2$. Create 1000 separate populations that each include 3 generations and use a histogram to visualize the empirical distribution of the number of offspring in the third generation. Estimate the expected number of 3rd generation offspring using the average across all 1000 simulations and estimate the variance of the number using the R var() function (we will learn the basis for these estimates in Chapter 8). Compare these estimates with the values computed according to the results in part a) and b).

4. (Optional - Based on Rice 4.81) Find the moment-generating function of a Bernoulli RV and use it to find the mean, variance and third central moment.

5. (Based on Rice 5.1) Let $X_1, X_2, ...$ be a sequence of independent random variables with $E[X_i] = \mu$ and $Var(X_i) = \sigma_i^2$. Show that if $n^{-2} \sum_{i=1}^{n} \sigma_i^2 \to 0$, then $\bar{X} \to \mu$ in probability.

6. (Optional - Based on Rice 5.5) Using moment-generating functions, show that as $n \to \infty, p \to 0$, and $np \to \lambda$, the binomial distribution with parameters n and p tends to the Poisson distribution.

7. (Based on Rice 5.16) Suppose that $X_1, ..., X_{20}$ are independent random variables with density functions $f(x) = 3x^2, 0 \leq x \leq 1$. Let $S = X_1 + ... + X_{20}$.

   (a) Use the central limit theorem to approximate $P(S \leq 14)$.

   (b) If you are instead asked to approximate $P(S \leq 15)$, what simplifications can be made in the calculation?

   (c) Validate the approximation by plotting the CLT-based density (compute this using dnorm()) and true density of S. Use the inverse CDF method to simulate from the true density and plot using a kernel density estimate (R code plot(kernel()), we'll learn the details of kernel density estimation later in the course).

8. (Based on Rice 5.21) We wish to evaluate the integral $I(f) = \int_a^b f(x)dx$ using a numerical estimate. Let g be a density function on $[a, b]$. Generate $X_1, ..., X_n$ from g and estimate I by $\hat{I}(f) = 1/n \sum_{i=1}^{n} f(X_i)/g(X_i)$.

   (a) Show that $E(\hat{I}(f)) = I(f)$

   (b) Demonstrate the result in a) via simulation with f(x) the density of the standard normal, a=0, b=1 and g(x) the density of the standard uniform distribution. Evaluate for $n = 5, ..., 100$. Plot $\hat{I}(f)$ as a function of n and include a horizontal line at $I(f)$.

   (c) (optional) Can this estimate be improved by choosing g to be other than uniform? Repeat the simulation in b) using a different choice of g (one you think will improve the estimate) and generate a new plot of $\hat{I}(f)$ vs. n that includes the estimates from both g functions. Discuss the relative estimation performance.