Charles Valentine

Homework 4

9/30/2016

Problem 1

a) Let $\bar{X}$ be the distribution of being independent random samples from a distribution with $\mu = 5$ and $\sigma = 3$ and $\sigma^2 = 9$. The sample mean of $\bar{X}$ is therefore $\frac{1}{5}\sum_{i=1}^{5}\mu_i = 5$ and the sample standard deviation $\left(\frac{1}{5}\right)^2 \sum_{i=1}^{5}\sigma_i^2 = \frac{\sigma^2}{5}$, $\mathrm{sd}(X) = \sqrt{\frac{\sigma^2}{5}} = \frac{\sigma}{\sqrt{5}} = 1.3416$ as predicted by the Central Limit Theorem.

b) You can find $P(2 < \bar{X} < 5.1) = P(\bar{X} < 5.1) - P(\bar{X} \leq 2)$ approximately by using the Central Limit Theorem to realize this probability over a normal distribution $N(\mu = 5, \sigma = \frac{3}{\sqrt{5}})$. See code.

OUTPUT:

```
P(2 < X̄ < 5.1) = P(X̄ < 5.1) − P(X̄ ≤ 2)) for N(μ = 5, σ = 3/√5) is:
[1] 0.7132
```

Problem 2

Let $\bar{Y}$ denote the average number of purines in $n = 100$ independent random microRNAs of size 20 with the probability of a purine in any location being binomially distributed with a probability $p = 0.7$. We are then looking for $P(\bar{Y} > 15)$. We begin with $X_1, \ldots, X_{100}$ as a random sample of size $n = 100$ each from the binomial distribution $\mathrm{binom}(size = 20, p = 0.7)$. The mean of this binomial distribution is $\mu = np = 14$ and the variance is $\sigma^2 = np(1-p) = 4.2$. We can then approximate $P(\bar{Y} > 15) = P(\bar{Y} = 20) - P(\bar{Y} = 15)$ by using the Central Limit Theorem:

$$\sigma_{\mathrm{normal}} = \frac{\sigma^2_{\mathrm{binomial}}}{\sqrt{nsims}} = 0.42$$

$$P(\bar{Y} > 15) = P(\bar{Y} = 20) - P(\bar{Y} = 15) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(20-\mu)^2/2\sigma^2} - \frac{1}{\sigma\sqrt{2\pi}} e^{-(15-\mu)^2/2\sigma^2}$$

$$P(\bar{Y} > 15) = \frac{1}{0.42\sqrt{2\pi}} e^{-(20-14)^2/2\cdot 0.42^2} - \frac{1}{0.42\sqrt{2\pi}} e^{-(15-14)^2/2\cdot 0.42^2} = 0.0086$$

Problem 3

Let $X(\mu = 9, \sigma^2 = 3)$ and $Y(\mu = 10, \sigma^2 = 5)$ and $Cov(X, Y) = 2$. The bivariate normal distribution of X and Y can be specified with the mean vector $(\mu_1, \mu_2) = (9, 10)$ and the covariance matrix as $\begin{bmatrix} {\sigma_1}^2 & \sigma_{12} \\ \sigma_{12} & {\sigma_2}^2 \end{bmatrix} = \begin{bmatrix} 3 & 2 \\ 2 & 5 \end{bmatrix}$. To estimate the probability $P(\bar{X} + 0.5 < \bar{Y})$ we will simulate 10 trials of the 50 independent measurements for each X and Y. We will then calculate the probability $P(\bar{X} + 0.5 < \bar{Y})$ and store that value. We will repeat this procedure 1000 times. The average of all the stored probabilities is the Monte Carlo approximation of the probability $P(\bar{X} + 0.5 < \bar{Y})$. We can assess the 95% confidence intervals to determine if this is an accurate approximation. See code for implementation.

OUTPUT:

```
P(X̄ + 0.5 < Ȳ) for (X1, Y2, ..., X50, Y50)
Sample mean:
[1] 29.9429

Sample 95% Confidence Intervals:
[1] 29.8764 30.0094
```

Problem 4

Please see code for implementation. The results show that the three independent random variables X_1, X_2 and X_3 as defined as $X_1 \sim \text{chisq}(df = 10)$, $X_2 \sim \text{Gamma}(\alpha = 1, \beta = 2)$, and $X_3 \sim \text{t}(m = 3)$ in the form $Y = \sqrt{X_1} X_2 + 4(X_3)^2$ have a mean of 13.5074 with the 95% confidence intervals at (13.3577, 13.6572). The simulation was run to represent 1,000 simulations of 10,000 realizations of each distribution. This simulation should offer a good approximation of the mean of Y and the tight confidence intervals support this.

OUTPUT:

```
Sample mean:
[1] 13.50744

Sample 95% Confidence Intervals:
[1] 13.35771 13.65717
```

Problem 5

The density function as described in Pevsner (2003, p.103) is defined as:

$$f(x) = (e^{-x})e^{-e^{-x}}$$

An extreme value distribution is setup following the routine of drawing 1000 numbers from a standard normal distribution and determining the maximum.

$$\text{maximas} = \max\left(Normal_i(n = 1000, \mu = 0, \sigma = 1)\right) \text{ for } i = 1, 2, \dots, 1000$$

Then we will subtract from these maximas $a(n)$ and divide by $b(n)$ where:

$$a(n) = \sqrt{2\log(n)} - \frac{1}{2}(\log(\log(\text{n})) + \log(4\pi)) \cdot (2\log(\text{n}))^{-0.5}$$

$$b(n) = (2\log(n))^{-0.5}$$

See code for implementation of simulation and plotting function. The figure below shows the approximated density of the normalized maximas. I found this question to be extremely confusing with many lexical errors that impaired my understanding of the end goal of this question. Not only did I learn nothing but I barely understood what the author intended for me when the text described adding functions (extreme value function, normal density function). Were we to add these functions to the plot or to the normalized maximas values? What would any of this achieve?