ANLY-511 Homework assigned on 09/16/15
Email a pdf file with solutions (preferred)
or hand it in at my office door by Wednesday, 9/23 11:59PM.
Five "short" and three "long" problems.

Explain your work and give concise reasoning. Attach `R` *code with comments if applicable. Do not print out any data or any detailed results of simulations. Your solutions for this homework set should fit on no more than six pages including graphs. In problems requiring graphs, you are allowed to just give the code to make the graphs and to describe them, without actually including them.*

9. (2 points) Suppose X has a Gamma distribution with shape parameter $r = 2.5$ and scale parameter $\rho = 5$. Use `R` to compute the following quantities: $Prob(X \leq 10)$, $Prob(X > 5$, $Prob(|X - 8| < 3$, and z such that $Prob(X < z) = .1$.

10. (2 points) Plot the cumulative distribution functions of a binomial distribution, $B(20, 1/3)$, and of a hypergeometric distribution with parameters $k = 20, M = 40$, $M + N = 120$ in the same figure. Use a staircase plot (type = 's'). These two distributions are supposed to be close. Can you confirm that?

11. (2 points) Plot the cumulative distribution function of a binomial distribution, $B(40, .3)$, and of a normal distribution, $N(12, 2.9)$ in the same figure, using a staircase book. These two distributions are supposed be close. Is that approximately right? What are the main differences between the two distributions?

12. (2 points) A graphical technique for checking whether a sample has an approximate normal distribution is a "quantile-quantile" plot. The `R` command is `qqnorm(x)`, where x is the vector containing the sample values. If the plot is approximately a straight line, then this suggests that the sample comes from a normal distribution. Practice this by making `qqnorm` plots of samples of size 10, 20, 40, 100, 1000 from a standard normal distribution. How close to a straight line are the plots in each case?Where in the plot are the main deviations from being a straight line?

13. (2 points) Refer to the `qqnorm` plot as explained in the previous problem. Make `qqnorm` plots of random samples from a $B(40, .3)$ distribution, for sample sizes between 20 and 1000. How their plots differ from straight lines? Can you explain why?

14. (5 points) Exercise 3.1 in Dalgaard.

15. (5 points) If X has a continuous distribution with cumulative distribution function F, then the new random variable $U = F(X)$ has a uniform $U(0, 1)$ distribution. Verify this with simulations for three different continuous distributions of your choice, by making a random sample of sufficient size, sorting it, plugging it into the cdf F, and plotting the result. Then prove it using algebra for the case where X has an exponential distribution with arbitrary parameter λ.

16. (5 points) Suppose $X = X_1 + X_2$ is the sum of two exponentially distributed random variables with the same parameter λ. Then X^α is very nearly normally distributed for a suitable choice of α. Determine an approximate value for α, using a simulation and `qqnorm` plots.

ANLY-511 Homework assigned on 09/04/15
Email a pdf file with solutions (preferred)
or hand it in at my office door by Wednesday, 9/16.
Five "short" problems and three "long" problems.

Explain your work and give concise reasoning. Attach `R` *code with comments if applicable. Do not print out any data or any detailed results of simulations. Your solutions for this homework set should fit on no more than four pages.*

1. (2 points) Let a be the 10th digit to the right of the decimal point of $\sin 1.23$. Let $b = \sqrt{a^2 + a^3}$. Let c be the number of digits to the left of the decimal point of $e^{(\ln b)^3}$. Let $d = \sum_{j=1}^{c} j^3$. Compute a, b, c, d using `R` wherever it is appropriate.

2. (2 points) Given p between 0 and 1, the `R` commands `myattempts(p)` and `1 + rgeom(1,p)` simulate exactly the same thing. Compare the times for using these functions, for at least five different values of p, including values close to 0 and close to 1. Summarize and explain what you see. Your solution should also include the `R` code for a single timing run.

3. (2 points) Consider the random variable defined by counting the number of tries until the first success, for independent trials with success probability p. This random variable has a geometric distribution (see course slides) and we can simulate it with the function `myattempts()` or using `rgeom()`, see the previous problem. The standard deviation of such a geometric distribution is known to equal $\frac{\sqrt{1-p}}{p}$. Verify this for at least five different values of p, with a suitable simulation. Report your `R` code with comments and give the numerical results. The `R` code for computing the standard deviation of a vector is `sd`.

4. (2 points) Problem 1.5 in ch. 1 of **Dalgaard**.

5. (2 points) In the statistical computing community, there is an ongoing debate on the comparison of `R` or SAS. Much of this debate is happening on the Internet. Look up some arguments for and against both SAS and `R` and summarize them in no more than half a page.

6. (5 points) Problem 6 in ch. 1 of **Chihara / Hesterberg**. *Read the chapter and review basic probability. Use* `R` *to calculate all probabilities. For part b, first compute the probability that you will not be in any single sample. For part c, use trial and error. It is sufficient to give the answer as a multiple of 100,000.*

7. (5 points) For a random variable X with a geometric distribution (as in problem 2 and 3) it is known that the expected value of its square is

$$\mathcal{E}(X^2) = \frac{2 - 3p + p^2}{(1-p)p^2}\,.$$

(You don't have to show this). This equation can be solved for p, and the result is

$$p = \frac{\sqrt{1 + 8\mathcal{E}(X^2)} - 1}{2\mathcal{E}(X^2)}$$

(You don't have to show this either). Since we can use data to estimate $\mathcal{E}(X^2)$, this suggests another plug-in estimator for p. Write down a formula for this plug-in estimator. Then use a simulation to assess its bias, if $p = .3$ and you are given

samples of size $n = 4$ from a geometric distribution. Include commented `R` code in your solution.

8. (5 points) Consider the following random experiment: draw a uniformly distributed random number X_1 from the interval $(0, 1)$. Next, draw a uniformly distributed random number X_2 from the interval $(0, X_1)$, a uniformly distributed random number X_3 from the interval $(0, X_2)$ and so on until X_6. What is the expected value of X_6? Use a simulation to give an approximate answer. Report your commented `R` code. *The `R` command for drawing a uniformly distributed random number from the interval* $(0, b)$ *is* `runif(1,max = b)`.

Bonus (3 points) Refer to problem 8. What is the exact expected value of X_6? What is the exact expected value of X_4? Justify your answer.