18.443 Problem Set 5 (Optional)
Statistics for Applications
Due Date: 3/20/2015
prior to 3:00pm

1. **Log Normal Distribution:** A random variable X follows a $Lognormal(\theta, \sigma^2)$ distribution if $Y = \ln(X)$ follows a $Normal(\theta, \sigma^2)$ distribution.

   For the normal random variable $Y = \ln(X)$

   - The probability density function of Y is
   $$f(y \mid \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y-\theta)^2}{\sigma^2}}, \quad -\infty < y < \infty.$$
   - The moment-generating function of Y is
   $$M_Y(t) = E[e^{tY} \mid \theta, \sigma^2] = e^{t\theta + \frac{1}{2}\sigma^2 t^2}$$

   (a). Compute the first two moments of a random variable $X \sim Lognormal(\theta, \sigma^2)$.

   $$\mu_1 = E[X \mid \theta, \sigma^2] \text{ and } \mu_2 = E[X^2 \mid \theta, \sigma^2]$$

   Hint: Note that $X = e^Y$ and $X^2 = e^{2Y}$ where $Y \sim N(\theta, \sigma^2)$ and use the moment-generating function of Y.

   (b). Suppose that $X_1, \ldots, X_n$ is an i.i.d. sample from the $Lognormal(\theta, \sigma^2)$ distribution of size n. Find the method of moments estimates of θ and σ^2.

   Hint: evaluate μ_2/μ_1^2 and find a method-of-moments estimate for σ^2 first.

   (c). For the log-normal random variable $X = e^Y$, where

   $$Y \sim Normal(\theta, \sigma^2),$$

   prove that the probability density of X is

   $$f(x \mid \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} (\frac{1}{x}) e^{-\frac{1}{2}\frac{(\ln(x)-\theta)^2}{\sigma^2}}, \quad 0 < x < \infty.$$

   (d). Suppose that $X_1, \ldots, X_n$ is an i.i.d. sample from the $Lognormal(\theta, \sigma^2)$ distribution of size n. Find the mle for θ assuming that σ^2 is known to equal σ_0^2.

   (e). Find the asymptotic variance of the mle for θ in (d).

2. The Pareto distribution is used in economics to model values exceeding a threshhold (e.g., liability losses greater than \$100 million for a consumer products company). For a fixed, known threshhold value of $x_0 > 0$, the density function is

$$f(x \mid x_0, \theta) = \theta x_0^{\theta} x^{-\theta-1}, \quad x \geq x_0, \text{ and } \theta > 1.$$

Note that the cumulative distribution function of X is

$$P(X \leq x) = F_X(x) = 1 - \left(\frac{x}{x_0}\right)^{-\theta}.$$

(a). Find the method-of-moments estimate of θ.

(b). Find the mle of θ.

(c). Find the asymptotic variance of the mle.

(d). What is the large-sample asymptotic distribution of the mle?

3. Distributions derived from Normal random variables. Consider two independent random samples from two normal distributions:

- $X_1, \ldots, X_n$ are n i.i.d. $Normal(\mu_1, \sigma_1^2)$ random variables.
- $Y_1, \ldots, Y_m$ are m i.i.d. $Normal(\mu_2, \sigma_2^2)$ random variables.

(a). If $\mu_1 = \mu_2 = 0$, find two statistics

$$T_1(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$$
$$T_2(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$$

each of which is a t random variable and which are statistically independent. Explain in detail why your answers have a t distribution and why they are independent.

(b). If $\sigma_1^2 = \sigma_2^2 > 0$, define a statistic

$$T_3(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$$

which has an F distribution.

An F distribution is determined by the numerator and denominator degrees of freedom. State the degrees of freedom for your statistic T_3.

(c). For your answer in (b), define the statistic

$$T_4(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = \frac{1}{T_3(X_1, \ldots, X_n, Y_1, \ldots, Y_m)}$$

What is the distribution of T_4 under the conditions of (b)?

(d). Suppose that $\sigma_1^2 = \sigma_2^2$. If $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, and $S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2$, are the sample variances of the two samples, show how to use the F distribution to find

$$P({S_X}^2/S_Y^2 > c).$$

(e). Repeat question (d) if it is known that $\sigma_1^2 = 2\sigma_2^2$.

4. **Hardy-Weinberg (Multinomial) Model of Gene Frequencies**

For a certain population, gene frequencies are in equilibrium: the genotypes AA, Aa, and aa occur with probabilities $(1-\theta)^2$, $2\theta(1-\theta)$, and θ^2. A random sample of 50 people from the population yielded the following data:

Genotype Type		
AA	Aa	aa
35	10	5

The table counts can be modeled as the multinomial distribution:

$$(X_1, X_2, X_3) \sim Multinomial(n = 50, p = ((1-\theta)^2, 2\theta(1-\theta), \theta^2).$$

(a). Find the mle of θ

(b). Find the asymptotic variance of the mle.

(c). What is the large sample asymptotic distribution of the mle?

(d). Find an approximate 90% confidence interval for θ. To construct the interval you may use the follow table of cumulative probabilities for a standard normal $N(0,1)$ random variable Z

$P(Z < z)$	z
0.99	2.326
0.975	1.960
0.950	1.645
0.90	1.182

(e). Using the mle $\hat{\theta}$ in (a), 1000 samples from the

$$Multinomial(n = 50, p = ((1-\hat{\theta})^2, 2\hat{\theta}(1-\hat{\theta}), \hat{\theta}^2))$$

distribution were randomly generated, and mle estimates were computed for each sample: $\hat{\theta}_j^*, j = 1, \ldots, 1000$.

For the true parameter θ_0, the sampling distribution of $\Delta = \hat{\theta} - \theta_0$ is approximated by that of $\tilde{\Delta} = \hat{\theta}^* - \hat{\theta}$. The 50-th largest value of $\tilde{\Delta}$ was $+0.065$ and the 50-th smallest value was -0.067.

Use this information and the estimate in (a) to construct a (parametric) bootstrap confidence interval for the true θ_0. What is the confidence level of the interval? (If you do not have an answer to part (a), assume the mle $\hat{\theta} = 0.25$).